On iterative refinement for the spectral decomposition of symmetric matrices

Alexander N. Malyshev

UNITÉ DE RECHERCHE
INRIA-RENNES

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél.: (1) 39 63 55 11

Rapports de Recherche

1992

25ème anniversaire

N° 1651

Programme 1
Architectures parallèles, Bases de données,
Réseaux et Systèmes distribués

ON ITERATIVE REFINEMENT FOR THE SPECTRAL DECOMPOSITION OF SYMMETRIC MATRICES

Alexander N. MALYSHEV

Mars 1992

RR_1651

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTEMES ALEATOIRES

Campus Universitaire de Beaulieu
35042 - RENNES CEDEX FRANCE
Tél. : 99 84 71 00 - Télex : UNIRISA 950 473 F
Télécopie : 99 38 38 32

On iterative refinement for the spectral decomposition of symmetric matrices

Raffinement itératif d'une décomposition spectrale de matrice symétrique [1]

Alexander N. Malyshev

Institute of Mathematics
Academy of Sciences
Universitetsky pr., 4
Novosibirsk, 630090
RUSSIA

Programme 6

Publication interne n^0 628 - 26 pages.

Janvier 1992

Abstract

A method is studied for computation of the spectral decomposition of a symmetric matrix with high precision. The approach used is based upon careful calculation of a solution of the Riccati equations arising in the refinement process. Detailed rounding error analysis and estimates of convergence rates are carried out.

Keywords : Symmetric eigenvalue problem, Riccati equation, iterative refinement, rounding error analysis.

Résumé

On développe ici une méthode pour diagonaliser une matrice symétrique avec grande précision. Le cœur de l'approche repose sur le calcul de la solution d'une équation de Riccati qui apparait dans le processus de raffinement. L'analyse complète des erreurs d'arrondi et une estimation des taux de convergence sont proposées.

Mots clés: Problème symétrique aux valeurs propres, équation de Riccati, raffinement itératif, erreur d'arrondi.

AMS (MOS) : 65F15.

[1]This work was carried out in the period June - November 1991, while the author was an ERCIM fellow at IRISA/INRIA, FRANCE. The visit was supported by the French Ministry of Research.
Ce travail a été mené par l'auteur pendant son séjour comme chercheur ERCIM à l'IRISA/INRIA. La visite était financée par le Ministère français de la Recherche et de la Technologie.

Introduction

The eigenvalue problem for a real dense symmetric $N \times N$-matrix A with small N can be efficiently solved by QR or the Sturm sequences methods. The latter method is exposed in [6], [9] with an exhaustive error analysis including the rounding errors.

The commonly used procedure consists of the following steps

a) tridiagonalization of A by means of orthogonal similarity transformations, usually, by the Householder transformations;

b) deflation of the computed symmetric tridiagonal matrix by the Jacobi-Givens rotation chains.

However, the resulting spectral decomposition of A, $X\Lambda X^{-1}$, can considerably differ from A. Here Λ is a diagonal matrix and X is nearly orthogonal. More precisely, one can obtain the following bounds:

$$\|AX - X\Lambda\| \leq f(N)\epsilon\|A\|, \tag{1}$$
$$\|X^TX - I\| \leq f(N)\epsilon,$$

with the relative precision ϵ of floating point system used and the polynomial $f(N)$ of small degree. For example, $f(N) \approx 12N^{5/2}$ in [6].

It is well known that if the error in the matrix A is of order $\delta\|A\|$ then due to the Courant-Fisher principle the error in the eigenvalues of A is also of order $\delta\|A\|$. When $\delta \approx \epsilon$, for example, then $\delta \ll f(N)\epsilon$ that is the spectral decomposition (1) can be deficient.

The question arises whether one can exploit the roughly computed decomposition (1) in order to calculate more accurate spectral decomposition. This could be an alternative to computation of (1) with double precision entirely.

In this paper we assess some ideas proposed in [3], [8] and [10]. Our method consists of the following basic steps:

1. reorthonormalization of the columns of X with higher precision;
2. calculation of $\hat{X}^TA\hat{X}$ with higher precision, where $\hat{X}$ is the result of the orthonormalization;
3. block diagonalization of $\hat{X}^TA\hat{X}$ by means of special Riccati equations solutions;
4. diagonalization of the diagonal blocks computed in step 3 if it is possible.

During the steps some tricks are used in order to make as less as possible of computations with higher precision.

This refinement procedure is provided with the complete theory of rounding errors. We obtained a priori error bounds for all stages of the process. Perhaps, a posteriori estimates sometimes will be sharper but by additional price. And, more important, a priori estimates enable to prove rigorously that the method always works.

In the paper we have not actually presented a strict scheme of the algorithm which is suitable for immediate implementation. Instead a few possible techniques are discussed in detail. Those who intend to implement this method can choose, for example, between orthonormalization processes, between amounts of work with double precision but the principal idea to exploit solutions of the Riccati equations always remains the same. Also choice of some critical parameters, like τ in section 3, is not strictly limited and left for additional implementation study.

1 Some technical lemmas

Proof of the following lemma is evident but the lemma is extremely powerful in manipulations with error estimates.

Lemma 1 *If* $|\alpha| < 1$ *then*

$$|1+\alpha| \le \frac{1}{1-|\alpha|}, \qquad \frac{1-|\alpha|}{1-|\alpha|/2} \le \sqrt{1+\alpha} \le 1+\frac{|\alpha|}{2}.$$

Let A be a symmetric $N \times N$-matrix, Λ be a symmetric $k \times k$-matrix ($k \le N$) and X be a $N \times k$-matrix such that

$$\|AX - X\Lambda\| \le \delta_1 \|A\| + \delta_0, \tag{2}$$

$$\|X^T X - I\| \le \delta_2. \tag{3}$$

Lemma 2 *If* $\delta_2 < 1$ *then from (2),(3) it follows that*

$$\|\Lambda\| \le \frac{1+\delta_1}{1-\delta_2}\|A\| + \frac{\delta_0}{1-\delta_2}. \tag{4}$$

PROOF: Since $\Lambda = (X^TX)^{-1}X^TAX - (X^TX)^{-1}X^T(AX - X\Lambda)$, $\|X\| \le \sqrt{1+\delta_2}$, $\|(X^TX)^{-1}X^T\| = \|(X^TX)^{-1/2}\| \le 1/\sqrt{1-\delta_2}$ then it holds that

$$\|\Lambda\| \le \sqrt{\frac{1+\delta_2}{1-\delta_2}}\|A\| + \|AX - X\Lambda\|\frac{1}{\sqrt{1-\delta_2}} \le$$

$$\le (1 + \frac{\delta_2}{1-\delta_2})\|A\| + \frac{\delta_1\|A\| + \delta_0}{1-\delta_2} = \frac{(1+\delta_1)\|A\| + \delta_0}{1-\delta_2}.$$

■

Lemma 3 *If* $k = N$ *and* $\delta_1 + \delta_2 < 1$ *then (2),(3) imply*

$$\|A\| \le \frac{\|\Lambda\|}{1-\delta_1-\delta_2} + \frac{\delta_0}{1-\delta_2}. \tag{5}$$

PROOF: From the equality $A = X\Lambda X^{-1} + (AX - X\Lambda)X^{-1}$ one can derive that

$$\|A\| \le \|\Lambda\|\|X\|\|X^{-1}\| + \|AX - X\Lambda\|\|X^{-1}\| \le \|\Lambda\|\sqrt{\frac{1+\delta_2}{1-\delta_2}} + \frac{\delta_1\|A\| + \delta_0}{\sqrt{1-\delta_2}},$$

$$\|A\| \le \|\Lambda\|\sqrt{\frac{1+\delta_2}{1-\delta_2}}/(1 - \frac{\delta_1}{\sqrt{1-\delta_2}}) + \frac{\delta_0}{\sqrt{1-\delta_2}} \le \|\Lambda\|\frac{1}{1-\delta_2}/(1 - \frac{\delta_1}{1-\delta_2}) + \frac{\delta_0}{1-\delta_2}.$$

■

Now we shall obtain the estimate which reveals relations between the matrices A and Λ in (2),(3). Introduce for convenience the notations

$$AX - X\Lambda = \Delta_1, \quad X^TX - I = \Delta_2.$$

Consider the matrix $\hat{X} = X(I+\Delta_2)^{-1/2}$ which obeys the identity $\hat{X}^T\hat{X} = I$. Having an estimate for the residual $\|A\hat{X} - \hat{X}\Lambda\|$ one can apply the Kahan's theorem which says that there exist k eigenvalues of the matrix A, $\mu_1 \ge \mu_2 \ge \ldots \ge \mu_k$, such that $|\lambda_i - \mu_i| \le \|A\hat{X} - \hat{X}\Lambda\|$, where $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_k$ are the eigenvalues of Λ. We have

$$A\hat{X} - \hat{X}\Lambda = AX - X\Lambda + AX[(I+\Delta_2)^{-1/2} - I] - X[(I+\Delta_2)^{-1/2} - I]\Lambda,$$

$$\|X[(I+\Delta_2)^{-1/2}-I]\| = \min_{|x^2-1|\le\delta_2} ||x|-1| \le \sqrt{1+\delta_2}-1 \le \delta_2/2,$$

$$\|A\hat{X}-\hat{X}\Lambda\| \le \delta_1\|A\| + \delta_0 + \frac{\delta_2}{2}(\|A\|+\|\Lambda\|) \le \delta_1\|A\| + \delta_0 + \frac{\delta_2}{2}\left[\|A\|\left(1+\frac{1+\delta_1}{1-\delta_2}\right)+\frac{\delta_0}{1-\delta_2}\right],$$

$$\|A\hat{X}-\hat{X}\Lambda\| \le \frac{\delta_1+\delta_2}{1-\delta_2}\|A\| + \frac{\delta_0}{1-\delta_2}. \tag{6}$$

Thus, the eigenvalues of the matrix Λ approximate k appropriate eigenvalues of the matrix A with the absolute error which is less than $(\delta_1+\delta_2)\|A\|/(1-\delta_2)+\delta_0/(1-\delta_2)$.

If, in addition, these k eigenvalues are well separated from other $N-k$ eigenvalues of A then the sharper enclosures hold [11].

2 Basic definitions for the rounding error analysis

Here we list estimates of rounding errors in arithmetic and matrix operations. Detailed discussions about the estimates may be found in [6], [9] and [4].

We assume that

$$(a \pm b)_{mach} = a(1+\alpha) \pm b(1+\beta) + \gamma,$$

where $|\alpha| \le \epsilon_1$, $|\beta| \le \epsilon_1$, $|\gamma| \le \epsilon_0$. Similarly,

$$(ab)_{mach} = ab(1+\alpha)+\gamma, \quad (a/b)_{mach} = (a/b)(1+\alpha)+\gamma,$$

$$\sqrt{a}_{mach} = \sqrt{a}(1+\alpha),$$

where $|\alpha| \le \epsilon_1$, $|\gamma| \le \epsilon_0$.

The constant ϵ_1 is a relative precision of floating point number system on a computer. The constant ϵ_0 equals to the underflow threshold, *underflow*, when using the arithmetic without denormalized numbers and equals to $\epsilon_1 \times$ *underflow* when using the arithmetic with the denormalized numbers.

For addition (subtraction) of $M \times N$-matrices it holds that

$$\|(A \pm B)_{mach} - (A \pm B)\|_F \le \epsilon_1\|A\|_F + \epsilon_1\|B\|_F + \epsilon_0\sqrt{MN}.$$

For product of a $M \times N$-matrix A by a scalar α we have the error bound

$$\|(\alpha A)_{mach} - \alpha A\|_F \le \epsilon_1\|\alpha A\|_F + \epsilon_0\sqrt{MN}.$$

For inner product of N-vectors one can prove the following estimate

$$|(x,y)_{mach} - (x,y)| \le \frac{N\epsilon_1}{1-(N-1)\epsilon_1/2}\|x\|\|y\| + \frac{(2N-1)\epsilon_0}{1-(N-1)\epsilon_1}.$$

The product of $M \times N$-matrix A and $N \times K$-matrix B satisfies the bound

$$\|(AB)_{mach} - (AB)\|_F \le \frac{N\epsilon_1}{1-(N-1)\epsilon_1/2}\|A\|_F\|B\|_F + \frac{(2N-1)\epsilon_0}{1-(N-1)\epsilon_1}\sqrt{MK}.$$

For calculation of residuals higher precision is often exploited, as usual, it is double precision. In the case $\epsilon_1^{(2)}$ and $\epsilon_0^{(2)}$ are utilized instead of ϵ_1 and ϵ_0. One can assume that $\epsilon_1^{(2)} \approx \epsilon_1^2$ and $\epsilon_0^{(2)} \approx \epsilon_0$.

When calculating with higher precision the operation of rounding to the lower precision is used. In the case we have the estimate

$$\|round(A) - A\|_F \le \epsilon_1\|A\|_F + \epsilon_0\sqrt{MN}.$$

3 General error analysis

We start with the inequalities (2),(3) for $k = N$:

$$AX - X\Lambda = \Delta_1, \quad \|\Delta_1\| \le \delta_1\|A\| + \delta_0,$$

$$X^TX - I = \Delta_2, \quad \|\Delta_2\| \le \delta_2.$$

It is suggested to compute the matrix $X(I - \frac{1}{2}\Delta_2)$ in order to approximate the matrix $X(I + \Delta_2)^{-1/2}$. So, let $Y = X(I - \frac{1}{2}\Delta_2) + E_Y$ with small $\delta_Y = \|E_Y\|$. By means of SVD (the singular value decomposition) of the matrix X one can show that $\|X(I - \frac{1}{2}\Delta_2)\| \le 1$ and

$$\|Y^TY - I\| \le \|(I - \frac{1}{2}\Delta_2)X^TX(I - \frac{1}{2}\Delta_2) - I\| + 2\|E_Y\|\|X(I - \frac{1}{2}\Delta_2)\| + \|E_Y\|^2 \le$$

$$\le \max_{|x^2-1|\le\delta_2} |x^2(1 - \frac{x^2-1}{2})^2 - 1| + 2\delta_Y + \delta_Y^2 \le$$

$$\le \max_{|x^2-1|\le\delta_2} |(x^2-1)^2(\frac{3}{4} - \frac{x^2-1}{4})| + 2\delta_Y + \delta_Y^2 \le \delta_2^2\frac{3+\delta_2}{4} + 2\delta_Y + \delta_Y^2 \le$$

$$\le \delta_3 = \frac{3\delta_2^2/4 + 2\delta_Y}{1 - \delta_2/3 - \delta_Y/2}. \tag{7}$$

Assume that the eigenvalue problem for the matrix Y^TAY can be solved with higher precision that is

$$\|A_1X_1 - X_1\Lambda_1\| \le \sigma_1\|A_1\| + \sigma_0, \tag{8}$$

$$\|X_1^TX_1 - I\| \le \sigma_2, \tag{9}$$

where $A_1 = Y^TAY + E_{A_1}$, $E_{A_1}^T = E_{A_1}$, $\|E_{A_1}\| \le \alpha_1\|A\| + \alpha_0$ is small, $\Lambda_1^T = \Lambda_1$ is diagonal, the matrix X_1 may have smaller number of columns than the number of rows. Assume also that $Z = YX_1 + E_Z$ with small $\delta_Z = \|E_Z\|$.

We wish to estimate the residuals $\|AZ - Z\Lambda_1\|$ and $\|Z^TZ - I\|$. At first,

$$Z^TZ - I = X_1^TY^TYX_1 - I + E_Z^TYX_1 + X_1^TY^TE_Z + E_Z^TE_Z =$$

$$= X_1^T(Y^TY - I)X_1 + (X_1^TX_1 - I) + E_Z^TYX_1 + X_1^TY^TE_Z + E_Z^TE_Z,$$

$$\|Z^TZ - I\| \le \sigma_2 + \delta_3\|X_1\|^2 + 2\|E_Z\|\|Y\|\|X_1\| + \|E_Z\|^2 \le$$

$$\le \sigma_2 + \delta_3(1 + \sigma_2) + 2\delta_Z(1 + \delta_Y)(1 + \sigma_2/2) + \delta_Z^2 \le$$

$$\le \frac{\sigma_2 + \delta_3 + 2\delta_Z}{1 - \sigma_2 - \delta_Z/2 - \delta_Y} \le \frac{\sigma_2 + 3\delta_2^2/4 + 2\delta_Y + 2\delta_Z}{1 - \sigma_2 - \delta_Z/2 - 3\delta_Y/2 - \delta_2/3}. \tag{10}$$

Since $(Y^TAY + E_{A_1})X_1 - X_1\Lambda_1 = R_1$ with $\|R_1\| \le \sigma_1\|A_1\| + \sigma_0$, then

$$AZ - Z\Lambda_1 = AYX_1 + AE_Z - YX_1\Lambda_1 - E_Z\Lambda_1 =$$

$$= Y(R_1 - E_{A_1}X_1) + (I - YY^T)AYX_1 + AE_Z - E_Z\Lambda_1,$$

$$\|AZ - Z\Lambda_1\| \le \|Y\|(\|R_1\| + \|E_{A_1}\|\|X_1\|) + \|I - YY^T\|\|A\|\|Y\|\|X_1\| + (\|A\| + \|\Lambda_1\|)\|E_Z\| \le$$

$$\le \|Y\|[\sigma_1\|A_1\| + \sigma_0 + (\alpha_1\|A\| + \alpha_0)\sqrt{1+\sigma_2} + \delta_3\|A\|\sqrt{1+\sigma_2}] + \delta_Z(\|A\| + \|\Lambda_1\|).$$

Owing to the estimates $\|Y\| \le 1+\delta_Y$, $\|A_1\| \le \|A\|/(1-\alpha_1-2\delta_Y)+\alpha_0$, $\|\Lambda_1\| \le (1+\sigma_1)\|A_1\|/(1-\sigma_2) + \sigma_0/(1-\sigma_2)$ one can obtain the estimate

$$\|AZ - Z\Lambda_1\| \le (1+\delta_Y)\left[\|A\|\left(\frac{\sigma_1}{1-\alpha_1-2\delta_Y} + \frac{\alpha_1+\delta_3}{1-\sigma_2}\right) + \left(\sigma_0 + \frac{\alpha_0}{1--\sigma_1-\sigma_2}\right)\right] +$$

$$+\delta_Z[\|A\|(1+\frac{1+\sigma_1}{1-\sigma_2-\alpha_1-2\delta_Y})+\frac{\sigma_0+\alpha_0}{1-\sigma_1-\sigma_2}]\leq$$

$$\leq\|A\|\frac{\sigma_1+\alpha_1+\delta_3+2\delta_Z}{1-\sigma_1/2-\sigma_2-\alpha_1-3\delta_Y}+\frac{\sigma_0+\alpha_0}{1-\sigma_1-\sigma_2-\delta_Y-\delta_Z}\leq$$

$$\leq\|A\|\frac{\sigma_1+\alpha_1+3\delta_2^2/4+2\delta_Y+2\delta_Z}{1-\sigma_1/2-\sigma_2-\alpha_1-7\delta_Y/2-\delta_2/3}+\frac{\sigma_0+\alpha_0}{1-\sigma_1-\sigma_2-\delta_Y-\delta_Z}. \tag{11}$$

As a corollary, the eigenvalues of the matrix Λ_1 are appropriate eigenvalues of A with the absolute error less than

$$errabs=\|A\|\frac{\sigma_1+\sigma_2+3\delta_2^2/2+\alpha_1+4\delta_Y+4\delta_Z}{1-\sigma_1/2-3\sigma_2-\alpha_1-3\delta_Z-7\delta_Y-2\delta_2/3-3\delta_2^2/4}+$$

$$+\frac{\sigma_0+\alpha_0}{1-\sigma_1-3\sigma_2-\delta_2/3-3\delta_2^2/4-9\delta_Y/2-7\delta_Z/2}\leq$$

$$\leq\|\Lambda\|\frac{\sigma_1+\sigma_2+3\delta_2^2/2+\alpha_1+4\delta_Y+4\delta_Z}{1-3\sigma_1/2-4\sigma_2-2\alpha_1-7\sigma_Z-11\delta_Y-2\delta_2/3-9\delta_2^2/4}+ \tag{12}$$

$$+\frac{\sigma_0+\alpha_0}{1-3\sigma_1/2-4\sigma_2-2\alpha_1-11\delta_Y-7\delta_Z-9\delta_2^2/4-2\delta_2/3}.$$

Finally, it remains to compare the matrices A_1 and Λ. In fact,

$$A_1=Y^TAY+E_{A_1}=[X(I-\Delta_2/2)+E_Y]^TA[X(I-\Delta_2/2)+E_Y]+E_{A_1}=$$

$$=(I-\Delta_2/2)X^TAX(I-\Delta_2/2)+(I-\Delta_2/2)X^TAE_Y+E_Y^TAX(I-\Delta_2/2)+E_Y^TAE_Y+E_{A_1}=$$

$$=(I-\Delta_2/2)X^T(X\Lambda+\Delta_1)(I-\Delta_2/2)+\ldots=(I-\Delta_2/2)(I+\Delta_2)\Lambda(I-\Delta_2/2)+$$

$$+(I-\Delta_2/2)X^T\Delta_1(I-\Delta_2/2)+(I-\Delta_2/2)X^TAE_Y+E_Y^TAX(I-\Delta_2/2)+E_Y^TAE_Y+E_{A_1},$$

$$\|A_1-\Lambda\|\leq\|(I-\Delta_2/2)(I+\Delta_2)\Lambda(I-\Delta_2/2)-\Lambda\|+$$

$$+\|\Delta_1\|\|I-\Delta_2/2\|+2\|A\|\|E_Y\|+\|A\|\|E_Y\|^2+\|E_{A_1}\|\leq$$

$$\leq\|\Lambda\|[(1+\max_{|x|\leq\delta_2}|\delta/2-\delta^2/2|)(1+\delta_2/2)-1]+(\delta_1\|A\|+\delta_0)(1+\delta_2/2)+2\|A\|\delta_Y+\|A\|\delta_Y^2+$$

$$+\alpha_1\|A_1\|+\alpha_0\leq\|\Lambda\|\left[(1+\frac{\delta_2}{2}(1+\delta_2))(1+\frac{\delta_2}{2})-1\right]+$$

$$+\|A\|\left[\delta_1(1+\delta_2/2)+2\delta_Y+\delta_Y^2+\alpha_1\right]+\alpha_0+\delta_0(1+\delta_2/2)\leq$$

$$\leq\|\Lambda\|\frac{\delta_2}{1-\delta_2}+\|A\|\frac{\delta_1+2\delta_Y+\alpha_1}{1-\delta_2/2-\delta_Y/2}+\frac{\delta_0+\alpha_0}{1-\delta_2/2},$$

$$\|A_1-\Lambda\|\leq\|A\|\frac{\delta_1+\delta_2+\alpha_1+2\delta_Y}{1-\delta_1-\delta_2-\delta_Y/2}+\frac{2\delta_0+\alpha_0}{1-2\delta_2}\leq \tag{13}$$

$$\leq\|\Lambda\|\frac{\delta_1+\delta_2+\alpha_1+2\delta_Y}{1-2\delta_1-2\delta_2-\delta_Y/2}+\frac{2\sigma_0+\alpha_0}{1-3\delta_1/2-5\delta_2/2-\alpha_1/2-3\delta_Y/2}.$$

Thus, the matrix A_1 is a perturbation of Λ of magnitude (13).

3.1 Calculation of δ_Y, δ_Z, α_1, α_0

1). In order to calculate the matrix $X(I-\frac{1}{2}\Delta_2) = X[I-\frac{1}{2}(X^TX-I)]$ the following procedure is suggested:

a) computation of X^TX with higher precision;
b) computation of $\Delta_2 = X^TX - I$ with higher precision;
c) rounding of Δ_2 to lower precision;
d) computation of $X\Delta_2$;
e) computation of $\frac{1}{2}X\Delta_2$;
f) computation of $X - \frac{1}{2}X\Delta_2$ with higher precision.

Then the standard analysis based upon the bounds from section 2 yields that

$$\text{a) } \|(X^TX)_{mach} - X^TX\|_F \le \frac{N\epsilon_1^{(2)}}{1-(N-1)\epsilon_1^{(2)}/2}\|X\|_F^2 + \frac{(2N-1)N\epsilon_0^{(2)}}{1-(N-1)\epsilon_1^{(2)}},$$

$$\text{b) } \|(\Delta_2)_{mach} - \Delta_2\|_F \le \epsilon_1^{(2)}\sqrt{N} + \frac{(N+1)\epsilon_1^{(2)}}{1-(N+1)\epsilon_1^{(2)}/2}\|X\|_F^2 + \frac{2N^2\epsilon_0^{(2)}}{1-N\epsilon_1^{(2)}},$$

$$\text{c) } \|(\Delta_2)_{mach} - \Delta_2\|_F \le \epsilon_1\|\Delta_2\|_F + \frac{\epsilon_1^{(2)}\sqrt{N}}{1-\epsilon_1} +$$

$$+\frac{(N+1)\epsilon_1^{(2)}}{1-\epsilon_1-(N+1)\epsilon_1^{(2)}/2}\|X\|_F^2 + \frac{2N^2\epsilon_0^{(2)}}{1-\epsilon_1-N\epsilon_1^{(2)}} + N\epsilon_0,$$

$$\text{d) } \|(X\Delta_2)_{mach} - X\Delta_2\|_F \le \frac{N\epsilon_1}{1-(N-1)\epsilon_1/2}\|X\|_F\|\Delta_2\|_F + \frac{\|X\|_F}{1-N\epsilon_1}(\epsilon_1\|\Delta_2\|_F +$$

$$+\frac{\epsilon_1^{(2)}\sqrt{N}}{1-\epsilon_1} + \frac{(N+1)\epsilon_1^{(2)}}{1-\epsilon_1-(N+1)\epsilon_1^{(2)}/2}\|X\|_F^2 + \frac{2N^2\epsilon_0^{(2)}+N\epsilon_0}{1-\epsilon_1-N\epsilon_1^{(2)}}\Big) + \frac{(2N-1)N\epsilon_0}{1-(N-1)\epsilon_1} \le$$

$$\le \frac{(N+1)\epsilon_1}{1-N\epsilon_1}\|X\|_F\|\Delta_2\|_F + \|X\|_F\frac{\epsilon_1^{(2)}\sqrt{N}+2N^2\epsilon_0^{(2)}+N\epsilon_0}{1-(N+1)\epsilon_1-N\epsilon_1^{(2)}} +$$

$$+\|X\|_F^3\frac{(N+1)\epsilon_1^{(2)}}{1-(N+1)\epsilon_1-(N+1)\epsilon_1^{(2)}/2} + \frac{(2N-1)N\epsilon_0}{1-(N-1)\epsilon_1},$$

$$\text{e) } \|(\frac{1}{2}X\Delta_2)_{mach} - \frac{1}{2}X\Delta_2\|_F \le \epsilon_0 N + \epsilon_1\|\frac{1}{2}X\Delta_2\|_F + \frac{1+\epsilon_1}{2}\Big[\frac{(N+1)\epsilon_1}{1-N\epsilon_1}\|X\|_F\|\Delta_2\|_F +$$

$$+\frac{\epsilon_1^{(2)}\sqrt{N}+2N^2\epsilon_0^{(2)}+N\epsilon_0}{1-(N+1)\epsilon_1-N\epsilon_1^{(2)}}\|X\|_F + \frac{(N+1)\epsilon_1^{(2)}}{1-(N+1)\epsilon_1-(N+1)\epsilon_1^{(2)}/2}\|X\|_F^3 +$$

$$+\frac{(2N-1)N\epsilon_0}{1-(N-1)\epsilon_1}\Big] \le \frac{(N+2)\epsilon_1}{1-(N+1)\epsilon_1}\frac{1}{2}\|X\|_F\|\Delta_2\|_F + \frac{\epsilon_1^{(2)}\sqrt{N}+2N^2\epsilon_0^{(2)}+N\epsilon_0}{1-(N+2)\epsilon_1-N\epsilon_1^{(2)}}\frac{1}{2}\|X\|_F +$$

$$+\frac{(N+1)\epsilon_1^{(2)}}{1-(N+2)\epsilon_1-(N+1)\epsilon_1^{(2)}/2}\frac{1}{2}\|X\|_F^3 + \frac{(N+1/2)N\epsilon_0}{1-N\epsilon_1},$$

$$\text{f) } \|Y_{mach} - Y\|_F \le \frac{1}{2}\|X\|_F\|\Delta_2\|_F\frac{(N+2)\epsilon_1+\epsilon_1^{(2)}}{1-(N+1)\epsilon_1-\epsilon_1^{(2)}} +$$

$$+\frac{1}{2}\|X\|_F\frac{\epsilon_1^{(2)}(\sqrt{N}+2)+2N^2\epsilon_0^{(2)}+N\epsilon_0}{1-(N+2)\epsilon_1-(N+1)\epsilon_1^{(2)}}+$$

$$+\frac{1}{2}\|X\|_F^3\frac{(N+1)\epsilon_1^{(2)}}{1-(N+2)\epsilon_1-(N+3)\epsilon_1^{(2)}/2}+\frac{(N+1/2)N\epsilon_0+N\epsilon_0^{(2)}}{1-N\epsilon_1-\epsilon_1^{(2)}}.$$

With the estimates $\|\Delta_2\|_F \le \delta_2\sqrt{N}$, $\|X\|_F \le \sqrt{1+\delta_2}\sqrt{N}$ we derive that

$$\|Y_{mach}-Y\|_F \le \frac{1}{2}\frac{N\delta_2[(N+2)\epsilon_1+\epsilon_1^{(2)}]}{1-\delta_2/2-(N+1)\epsilon_1-\epsilon_1^{(2)}}+\frac{1}{2}\frac{\sqrt{N}[\epsilon_1^{(2)}(\sqrt{N}+2)+2N^2\epsilon_0^{(2)}+N\epsilon_0]}{1-\delta_2/2-(N+2)\epsilon_1-(N+1)\epsilon_1^{(2)}}+$$

$$+\frac{1}{2}\frac{N^{3/2}(N+1)\epsilon_1^{(2)}}{1-3\delta_2/2-(N+2)\epsilon_1-(N+3)\epsilon_1^{(2)}/2}+\frac{(N+1/2)N\epsilon_0+N\epsilon_0^{(2)}}{1-N\epsilon_1-\epsilon_1^{(2)}},$$

$$\|Y_{mach}-Y\|_F \le \delta_Y = \frac{1}{2}\frac{N(N+2)\epsilon_1\delta_2}{1-3\delta_2/2-(N+2)\epsilon_1-(N+1)\epsilon_1^{(2)}}+$$

$$+\frac{1}{2}\frac{(N^{5/2}+N^{3/2}+N+2\sqrt{N}+N\delta_2)\epsilon_1^{(2)}+(2N^2+N^{3/2}+N)\epsilon_0+(2N^{5/2}+N)\epsilon_0^{(2)}}{1-3\delta_2/2-(N+2)\epsilon_1-(N+1)\epsilon_1^{(2)}}.$$

Notice that the matrix Y_{mach} is stored in higher precision.

2). The matrix $A_1 = Y^TAY$ is calculated by the following algorithm:

a) computation of AY with higher precision;

b) calculation of $Y\Lambda$ with higher precision;

c) calculation of $AY - Y\Lambda$ with higher precision and rounding of the result to lower precision;

d) calculation of $P = Y^T(AY - Y\Lambda)$;

e) the matrix Y^TAY is represented by $\Lambda + P$ but Λ and P are stored separately and only the lower triangular part of P is referenced.

For precise calculations it holds that

$$\|Y^TAY - [\Lambda + Y^T(AY - Y\Lambda)]\| = \|(Y^TY - I)\Lambda\| \le \delta_3\|\Lambda\|.$$

The rounding error analysis yields the following bounds:

$$\text{a) } \|(AY)_{mach} - AY\|_F \le \frac{N\epsilon_1^{(2)}}{1-(N-1)\epsilon_1^{(2)}/2}\|A\|_F\|Y\|_F + \frac{(2N-1)N\epsilon_0^{(2)}}{1-(N-1)\epsilon_1^{(2)}},$$

$$\text{b) } \|(Y\Lambda)_{mach} - Y\Lambda\|_F \le \epsilon_1^{(2)}\|Y\|_F\|\Lambda\| + \epsilon_0^{(2)}N,$$

$$\text{c) } \|(AY - Y\Lambda)_{mach}^{(2)} - (AY - Y\Lambda)\|_F \le \frac{(N+1)\epsilon_1^{(2)}}{1-(N+1)\epsilon - 1^{(2)}/2}\|A\|_F\|Y\|_F +$$

$$+\frac{2\epsilon_1^{(2)}}{1-\epsilon_1^{(2)}}\|Y\|_F\|\Lambda\| + \frac{2N^2\epsilon_0^{(2)}}{1-N\epsilon_1^{(2)}},$$

$$\|(AY - Y\Lambda)_{mach} - (AY - Y\Lambda)\|_F \le \epsilon_1\|AY - Y\Lambda\|_F +$$

$$+\frac{(N+1)\epsilon_1^{(2)}}{1-\epsilon_1-(N+1)\epsilon_1^{(2)}/2}\|A\|_F\|Y\|_F + \frac{2\epsilon_1^{(2)}}{1-\epsilon_1-\epsilon_1^{(2)}}\|Y\|_F\|\Lambda\| + \frac{2N^2\epsilon_0^{(2)}}{1-\epsilon_1-N\epsilon_1^{(2)}},$$

$$\text{d) } \|P_{mach} - \|_F \le \frac{N\epsilon_1}{1-(N-1)\epsilon_1/2}\|Y\|_F\|AY - Y\Lambda\|_F + \frac{(2N-1)N\epsilon_0}{1-(N-1)\epsilon_1} +$$

$$+\frac{\|Y\|_F}{1-N\epsilon_1}\left[\epsilon_1\|AY - Y\Lambda\|_F + \frac{(N+1)\epsilon_1^{(2)}}{1-\epsilon_1-(N+1)\epsilon_1^{(2)}/2}\|A\|_F\|Y\|_F + \right.$$

$$\left. +\frac{2\epsilon_1^{(2)}}{1-\epsilon_1-\epsilon_1^{(2)}}\|Y\|_F\|\Lambda\| + \frac{2N^2\epsilon_0^{(2)}}{1-\epsilon_1-N\epsilon_1^{(2)}}\right] \le$$

$$\le \frac{(N+1)\epsilon_1}{1-N\epsilon_1}\|Y\|_F\|AY - Y\lambda\|_F + \frac{(N+1)\epsilon_1^{(2)}}{1-(N-1)\epsilon_1-(N+1)\epsilon_1^{(2)}/2}\|A\|_F\|Y\|_F^2 +$$

$$+\frac{2\epsilon_1^{(2)}}{1-(N+1)\epsilon_1-\epsilon_1^{(2)}}\|\Lambda\|\|Y\|_F + \frac{2N^2\epsilon_0^{(2)}}{1-(N+1)\epsilon_1-N\epsilon_1^{(2)}}\|Y\|_F + \frac{(2N-1)N\epsilon_0}{1-(N-1)\epsilon_1}.$$

e) Let $\overline{P}_{mach}$ denote the symmetric matrix which is represented by the lower triangle of P_{mach}. Then

$$\|\overline{P}_{mach} - P_{mach}\|_F \le \sqrt{2}\|(Y^TY - I)\Lambda\|_F \le \sqrt{2N}\delta_3\|\Lambda\|,$$

$$\|\overline{P}_{mach} - Y^TAY\| \le (1+\sqrt{2N})\delta_3\|\Lambda\| + \|P_{mach} - P\|_F.$$

Since $\|Y\|_F \le \sqrt{N}(1+\delta_Y)$,

$$AY - Y\Lambda = AX(I - \Delta_2/2) - X(I - \Delta_2/2)\Lambda + AE_Y - E_Y\Lambda =$$

$$= \Delta_1 + AE_Y - E_Y\Lambda - (AX\Delta_2 - X\Delta_2\Lambda)/2,$$

$$\|AY - Y\Lambda\|_F \le (\delta_1 + \delta_Y + \frac{\delta_2}{2-\delta_2})\sqrt{N}\|A\| + (\delta_Y + \frac{\delta_2}{2-\delta_2}\sqrt{N}\|\Lambda\| + \delta_0\sqrt{N},$$

then

$$\|\overline{P}_{mach} - Y^TAY\| \le (1+\sqrt{2N})\delta_2\|\Lambda\| + \frac{(N+1)\epsilon_1}{1-N\epsilon_1}\sqrt{N}(1+\delta_Y)[(\delta_1+\delta_Y+\frac{\delta_2}{2-\delta_2})\sqrt{N}\|A\| +$$

$$+(\delta_Y + \frac{\delta_2}{2-\delta_2})\sqrt{N}\|\Lambda\| + \delta_0\sqrt{N}] + \frac{(N+1)\epsilon_1^{(2)}}{1-(N-1)\epsilon_1-(N+1)\epsilon_1^{(2)}/2}\|A\|N^{3/2}(1+\delta_Y)^2 +$$

$$+\frac{2\epsilon_1^{(2)}}{1-(N+1)\epsilon_1-\epsilon_1^{(2)}}\|\Lambda\|N(1+\delta_Y)^2 + \frac{2N^2\epsilon_0^{(2)}}{1-(N+1)\epsilon_1-N\epsilon_1^{(2)}}\sqrt{N}(1+\delta_Y) + \frac{(2N-1)N\epsilon_0}{1-(N-1)\epsilon_1} \le$$

$$\le \left[\frac{N(N+1)\epsilon_1(\delta_1+\delta_Y+\delta_2/2)}{1-N\epsilon_1-\delta_2/2-\delta_Y} + \frac{N^{3/2}(N+1)\epsilon_1^{(2)}}{1-(N-1)\epsilon_1-(N+1)\epsilon_1^{(2)}/2-2\delta_Y}\right]\|A\| +$$

$$+\left[(1+\sqrt{2N})\delta_3 + \frac{N(N+1)\epsilon_1(\delta_Y+\delta_2/2)}{1-N\epsilon_1-\delta_2/2-\delta_Y} + \frac{2N\epsilon_1^{(2)}}{1-(N+1)\epsilon_1-\epsilon_1^{(2)}-2\delta_Y}\right]\|\Lambda\| +$$

$$+\left[\frac{N(N+1)\epsilon_1\delta_0}{1-N\epsilon_1-\delta_Y} + \frac{2N^2\epsilon_0^{(2)}\sqrt{N}+(2N-1)N\epsilon_0}{1-(N+1)\epsilon_1-N\epsilon_1^{(2)}-\delta_Y}\right] \le \alpha_1\|A\| + \alpha_0$$

with

$$\alpha_1 = \frac{N(N+1)(\delta_1+\delta_2+2\delta_Y)\epsilon_1 + (1+\sqrt{2N}\delta_3 + (N^{3/2}(N+1)+2N)\epsilon_1^{(2)}}{1-(N+1)\epsilon_1-(N+1)\epsilon_1^{(2)}/2-\delta_1-3\delta_2/2-2\delta_Y},$$

$$\alpha_0 = \frac{[N(N+1)\epsilon_1 + (1+\sqrt{2N}\delta_3 + N(N+1)(\delta_Y+\delta_2/2)\epsilon_1) + 2N\epsilon_1^{(2)}]\delta_0}{1-(N+1)\epsilon_1-\epsilon_1^{(2)}-3\delta_2/2-2\delta_Y} +$$

$$+\frac{2N^2\sqrt{N}\epsilon_0^{(2)} + (2N-1)N\epsilon_0}{1-(N+1)\epsilon_1 - N\epsilon_1^{(2)} - \delta_Y}.$$

3). At last, when computing YX_1 with higher precision it holds that:

$$\|(YX_1)_{mach} - YX_1\|_F \le \frac{N\epsilon_1^{(2)}}{1-(N-1)\epsilon_1^{(2)}/2}\|Y\|_F\|X_1\|_F + \frac{(2N-1)N\epsilon_0^{(2)}}{1-(N-1)\epsilon_1^{(2)}} \le$$

$$\le \frac{N^2\epsilon_1^{(2)}}{1-\sigma_2-\delta_Y-(N-1)\epsilon_1^{(2)}/2} + \frac{(2N-1)N\epsilon_0^{(2)}}{1-(N-1)\epsilon_1^{(2)}} \le \delta_Z = \frac{N^2\epsilon_1^{(2)} + (2N-1)N\epsilon_0^{(2)}}{1-\sigma_2-\delta_Y-(N-1)\epsilon_1^{(2)}}.$$

4 Numerical solution of the Riccati equation

Consider the matrix Riccati equation of a particular type:

$$A_{22}R - RA_{11} + A_{21} - RA_{12}R = 0, \tag{14}$$

where $A_{11} = \Lambda_1 + \Delta_{11}$, $A_{22} = \Lambda_2 + \Delta_{22}$ with small $\|\Delta_{11}\|$, $\|\Delta_{22}\|$ and with the diagonal matrices Λ_1 and Λ_2. Moreover, suppose that either $\lambda_{max}(\Lambda_1) - \lambda_{min}(\Lambda_2) \le \tau$ or $\lambda_{max}(\Lambda_2) - \lambda_{min}(\Lambda_1) \le \tau$ with $\tau > \|\Delta_{11}\| + \|\Delta_{22}\|$. This means that the diagonal elements of Λ_1 are separated from the diagonals of Λ_2 by an interval of length τ.

Theorem 1 (Stewart G.W.) *If*

$$k = \frac{\|A_{12}\|\|A_{21}\|_F}{(\tau - \|\Delta_{11}\| - \|\Delta_{22}\|)^2} < \frac{1}{4} \tag{15}$$

then there exists a solution of the Riccati equation in the ball

$$\|R\|_F \le \frac{1-\sqrt{1-4k}}{2k}\frac{\|A_{21}\|_F}{\tau - \|\Delta_{11}\| - \|\Delta_{22}\|} < \frac{2\|A_{21}\|_F}{\tau - \|\Delta_{11}\| - \|\Delta_{22}\|}.$$

PROOF. Let us study iteration of the form

$$R_0 = 0, \quad R_{k+1} = -\mathcal{L}^{-1}(\Delta_{22}R_k - R_k\Delta_{11} + A_{21} - R_kA_{12}R_k), \tag{16}$$

where $\mathcal{L}$ is operator $\mathcal{L}X = \Lambda_2 X - X\Lambda_1$. It is evidently that

$$\|\mathcal{L}^{-1}\|_F = \sup_{\|X\|_F=1}\|\mathcal{L}^{-1}X\|_F = 1/\tau.$$

From (16) one can easily derive the inequality

$$\|R_{k+1}\|_F \le \frac{1}{\tau}[\|R_k\|_F(\|\Delta_{11}\| + \|\Delta_{22}\|) + \|A_{21}\|_F + \|a_{21}\|\|R_k\|_F^2].$$

Introducing the notations

$$\overline{a} = (\|\Delta_{11}\| + \|\Delta_{22}\|)/\tau, \ \overline{b} = \|A_{12}\|/\tau, \ \overline{c} = \|A_{21}\|_F/\tau,$$

we obtain the sequence x_k which satisfies the equation

$$x_0 = 0, \ x_{k+1} = \overline{a}x_k + \overline{b}x_k^2 + \overline{c},$$

and bounds $\|R_k\|_F$, $\|R_k\|_F \le x_k$.

Let us rewrite the equation $x_{k+1} = \overline{a}x_k + \overline{b}x_k^2 + \overline{c}$ in the form $x_{k+1} - x_k = (\overline{a} - 1)x_k + \overline{b}x_k^2 + \overline{c}$. Since $(\overline{a} - 1)x_k + \overline{b}x_k^2 + \overline{c} = \overline{b}(x_k - \overline{z}_1)(x_k - \overline{z}_2)$ with

$$\overline{z}_1 = \frac{1 - \overline{a} - \sqrt{(1-\overline{a})^2 - 4\overline{b}\overline{c}}}{2\overline{b}}, \quad \overline{z}_2 = \frac{1 - \overline{a} + \sqrt{(1-\overline{a})^2 - 4\overline{b}\overline{c}}}{2\overline{b}},$$

then $x_{k+1} - \overline{z}_1 = x_k - \overline{z}_1 + \overline{b}(x_k - \overline{z}_1)(x_k - \overline{z}_2) = (x_k - \overline{z}_1)[1 + \overline{b}(x_k - \overline{z}_2)]$. The condition $4\overline{b}\overline{c} \le (1-\overline{a})^2$ that is (15) yields $0 \le 1 - \overline{b}\overline{z}_2 \le 1 + \overline{b}(x_k - \overline{z}_2)$. Therefore, the sequence x_k belongs to the interval $[0, \overline{z}_1]$ and $x_k \to \overline{z}_1$ at least linearly with the coefficient $\rho = 1 + \overline{b}(\overline{z}_1 - \overline{z}_2) = 1 - \left(1 - \frac{\|\Delta_{11}\| + \|\Delta_{22}\|}{\tau}\right)\sqrt{1 - 4k}$. Thus, there is a solution of the Riccati equation in the ball $\|R\|_F \le \overline{z}_1 = \frac{1-\overline{a}}{2\overline{b}}\left(1 - \sqrt{1 - \frac{4\overline{b}\overline{c}}{(1-\overline{a})^2}}\right) = \frac{1-\sqrt{1-4k}}{2k}\frac{\overline{c}}{1-\overline{a}}$. ■

Now we shall study the process (16) when calculating with rounding errors. In the case the rounding errors can be simulated by the iteration

$$X_{k+1} = -\mathcal{L}^{-1}(\Delta_{22}X_k - X_k\Delta_{11} + A_{21} - X_kA_{12}X_k + \xi_1 + \xi_2) + \xi_3,$$

where

$$\|\xi_1\|_F \le \beta_1\|X_k\|_F + \beta_2\|X_k\|_F^2 + \beta_3,$$
$$\|\xi_2\|_F \le \gamma_1\|\Delta_{22}X_k - X_k\Delta_{11} + A_{21} - X_kA_{12}X_k + \xi_1\|_F,$$
$$\|\xi_3\|_F \le \gamma_2$$

with some β_1, β_2, β_3, γ_1 and γ_2 which will be evaluated later.

We have the estimate

$$\|X_{k+1}\|_F \le \gamma_2 + (1+\gamma_1)\|\mathcal{L}^{-1}\|_F[(\|\Delta_{11}\| + \|\Delta_{22}\| + \beta_1)\|X_k\|_F + (\|A_{12}\| + \beta_2)\|X_k\|_F^2 +$$
$$+(\|A_{21}\|_F + \beta_3)] \le a\|X_k\|_F + b\|X_k\|_F^2 + c$$

with

$$a = (1+\gamma_1)\frac{\|\Delta_{11}\| + \|\Delta_{22}\| + \beta_1}{\tau}, \quad b = (1+\gamma_1)\frac{\|A_{12}\| + \beta_2}{\tau}, \quad c = (1+\gamma_1)\frac{\|A_{21}\|_F + \beta_3}{\tau} + \gamma_2.$$

Repeating the proof of the theorem: if $4bc \le (1-a)^2$, i.e.

$$\frac{(\|A_{12}\| + \beta_2)[\|A_{21}\|_F + \beta_3 + \tau\gamma_2/(1+\gamma_1))}{[\tau/(1+\gamma_1) - \|\Delta_{11}\| - \|\Delta_{22}\| - \beta_1]^2} < \frac{1}{4},$$

and $\tau/(1+\gamma_1) - \|\Delta_{11}\| - \|\Delta_{22}\| - \beta_1 > 0$, then

$$\|X_k\|_F < \frac{2c}{1-a} = \frac{2[\|A_{21}\|_F + \beta_3 + \tau\gamma_2/(1+\gamma_1)]}{\tau/(1+\gamma_1) - \|\Delta_{11}\| - \|\Delta_{22}\| - \beta_1}.$$

Later we shall need an estimate for $\|X_k - X_{k-1}\|_F$. It holds that:

$$\|X_{k+1} - X_k\|_F \le 2\gamma_2 + \|\mathcal{L}^{-1}\|_F\{(\|\Delta_{11}\| + \|\Delta_{22}\|)\|X_k - X_{k-1}\|_F +$$
$$+2\|A_{12}\|\|X_{k-1}\|_F\|X_k - X_{k-1}\|_F + \|A_{12}\|\|X_k - X_{k-1}\|_F^2 +$$
$$+[(1+\gamma_1)\beta_1 + \gamma_1(\|\Delta_{11}\| + \|\Delta_{22}\|)](\|X_k\|_F + \|X_{k-1}\|_F) +$$
$$+[(1+\gamma_1)\beta_2 + \gamma_1\|A_{12}\|](\|X_k\|_F^2 + \|X_{k-1}\|_F^2) + [2(1+\gamma_1)\beta_3 + 2\gamma_1\|A_{21}\|_F]\} \le$$
$$\le a_1\|X_k - X_{k-1}\|_F + b_1\|X_k - X_{k-1}\|_F^2 + c_1$$

with

$$a_1 = \frac{\|\Delta_{11}\| + \|\Delta_{22}\|}{\tau} + \frac{2\|A_{12}\|}{\tau}\frac{2c}{1-a}, \quad b_1 = \frac{\|A_{12}\|}{\tau},$$

$$c_1 = 2\gamma_2 + [(1+\gamma_1)\beta_1 + \gamma_1(\|\Delta_{11}\| + \|\Delta_{22}\|)]\frac{4c}{1-a} + [(1+\gamma_1)\beta_2 + \gamma_1\|A_{12}\|]\frac{8c^2}{(1-a)^2} +$$

$$+[2(1+\gamma_1)\beta_3 + 2\gamma_1\|A_{21}\|_F],$$

$$\|X_1 - X_0\|_F = \|X_1\|_F \le \|A_{21}\|_F/\tau.$$

The sequence y_k satisfying

$$y_{k+1} = a_1 y_k + b_1 y_k^2 + c_1, \ \ y_1 = \|A_{21}\|_F/\tau,$$

is a bound for $\|X_k - X_{k-1}\|_F$. Under the conditions $\sqrt{4b_1c_1} \le (1-a_1)$ and $y_1 < (1 - a_1 + \sqrt{(1-a_1)^2 - 4b_1c_1})/(2b_1)$ one can show the convergence

$$y_k \to z = (1 - a_1 - \sqrt{(1-a_1)^2 - 4b_1c_1})/(2b_1), \ \ |z| < \frac{2c_1}{1-a_1}.$$

In particular, if $y_1 \le 1$, $2b_1 < 1 - a_1$ then

$$\|X_k - X_{k-1}\|_F \le y_k \le (a_1 + b_1)^k y_1 + \frac{c_1}{1 - a_1 - b_1}.$$

Our aim is to evaluate $r_k = \|A_{22}X_k - X_kA_{11} + A_{21} - X_kA_{12}X_k\|_F$. One obtains that

$$r_k = \|\Delta_{22}X_k - X_k\Delta_{11} + A_{21} - X_kA_{12}X_k - (\Delta_{22}X_{k-1} - X_{k-1}\Delta_{11} + A_{21} - X_{k-1}A_{12}X_{k-1} +$$

$$+\xi_1^{(k-1)} + \xi_2^{(k-1)}) + \mathcal{L}\xi_3\|_F \le (\|\Delta_{11}\| + \|\Delta_{22}\|)\|X_k - X_{k-1}\|_F + 2\|A_{12}\|\|X_{k-1}\|_F\|X_k - X_{k-1}\|_F +$$

$$+\|A_{12}\|\|X_k - X_{k-1}\|_F^2 + (1+\gamma_1)(\beta_1\|X_{k-1}\|_F + \beta_2\|X_{k-1}\|_F^2 + \beta_3) + \gamma_1(\|\Delta_1\| + \|\Delta_{22}\|)\|X_{k-1}\|_F +$$

$$+\gamma_1\|A_{21}\|_F + \gamma_1\|A_{12}\|\|X_{k-1}\|_F^2 + (\|\Delta_{11}\| + \|\Delta_{22}\|)\gamma_2 \le$$

$$\le (\|\Delta_{11}\| + \|\Delta_{22}\| + 2\|A_{12}\|\frac{2c}{1-a})\|X_k - X_{k-1}\|_F + \|A_{12}\|\|X_k - X_{k-1}\|_F^2 + [(1+\gamma_1)\beta_1 +$$

$$+\gamma_1(\|\Delta_{11} + \|\Delta_{22}\|)]\frac{2c}{1-a} + [(1+\gamma_1)\beta_2 + \gamma_1\|A_{12}\|]\frac{4c^2}{(1-a)^2} + [(1+\gamma_1)\beta_3 + \gamma_1\|A_{21}\|_F] +$$

$$+(\|\Delta_{11}\| + \|\Delta_{22}\|)\gamma_2 \le \tau[a_1y_k + b_1y_k^2 + c_1] - \tau c_1 + \frac{c_1 - 2\gamma_2}{2} + (\|\Delta_{11}\| + \|\Delta_{22}\|)\gamma_2 =$$

$$= \tau y_{k+1} + c_1(1/2 - \tau) + (\|\Delta_{11}\| + \|\Delta_{22}\| - 1)\gamma_2 \le \tau y_1(a_1 + b_1)^{k+1} + \frac{\tau c_1}{1 - a_1 - b_1} +$$

$$+c_1/2 - \tau c_1 + (\|\Delta_{11}\| + \|\Delta_{22}\| - 1)\gamma_2 \le$$

$$\le \tau y_1(a_1 + b_1)^{k+1} + c_1(1/2 + \tau\frac{a_1 + b_1}{1 - a_1 - b_1}) + (\|\Delta_{11}\| + \|\Delta_{22}\| - 1)\gamma_2.$$

Choosing k such that $\tau y_1(a_1 + b_1)^{k+1} < c_1/2$, i.e. $k + 1 \ge \log[c_1/(2\tau y_1)]/\log(a_1 + b_1)$ we obtain the ultimate estimate of the form

$$r_k \le c_1(1 + \tau\frac{a_1 + b_1}{1 - a_1 - b_1}) + (\|\Delta_{11}\| + \|\Delta_{22}\| - 1)\gamma_2.$$

Turn to evaluation of the coefficients β_1, β_2, β_3, γ_1, γ_2. The matrix ξ_1 is the matrix of errors when calculating $\Delta_{22}X_k - X_k\Delta_{11} + A_{21} - X_kA_{12}X_k$. The matrix X_k is a $M \times K$-matrix, $M + K = N$. It holds that

$$\|(\Delta_{22}X_k)_{mach} - \Delta_{22}X_k\|_F \le \frac{M\epsilon_1}{1 - (M-1)\epsilon_1/2}\|\Delta_{22}\|_F\|X_k\|_F + \frac{(2M-1)\sqrt{MK}}{1 - (M-1)\epsilon_1/2}\epsilon_0,$$

$$\|(X_k\Delta_{11})_{mach} - X_k\Delta_{11}\|_F \le \frac{K\epsilon_1}{1-(K-1)\epsilon_1/2}\|\Delta_{11}\|_F\|X_k\|_F + \frac{(2K-1)\sqrt{MK}}{1-(K-1)\epsilon_1/2}\epsilon_0,$$

$$\|(X_kA_{12})_{mach} - X_kA_{12}\|_F \le \frac{K\epsilon_1}{1-(K-1)\epsilon_1/2}\|X_k\|_F\|A_{12}\|_F + \frac{(2K-1)M}{1-(K-1)\epsilon_1/2}\epsilon_0,$$

$$\|(X_kA_{12}X_k)_{mach} - X_kA_{12}X_k\|_F \le \frac{\|X_k\|_F}{1-M\epsilon_1}\|(X_kA_{12})_{mach} - X_kA_{12}\|_F +$$

$$+\frac{M\epsilon_1}{1-(M-1)\epsilon_1/2}\|A_{12}\|\|X_k\|_F^2 + \frac{(2M-1)\sqrt{MK}}{1-(M-1)\epsilon_0/2} \le$$

$$\le \frac{(M+K)\epsilon_1}{1-(M+K)\epsilon_1}\|A_{12}\|_F\|X_k\|_F^2 + \frac{2KM\epsilon_0}{1-(M+K)\epsilon_1}\|X_k\|_F + \frac{(2N-1)\sqrt{MK}}{1-(N-1)\epsilon_1/2}\epsilon_0,$$

$$\|\xi_1\|_F \le [(1+\epsilon_1)^3 - 1][(\|\Delta_{11}\| + \|\Delta_{22}\|)\|X_k\|_F + \|A_{12}\|\|X_k\|_F^2 + \|A_{21}\|_F] +$$

$$+(1+\epsilon_1)^3\{[\frac{K\epsilon_1}{1-(K-1)\epsilon_1/2}\|\Delta_{11}\|_F + \frac{M\epsilon_1}{1-(M-1)\epsilon_1/2}\|\Delta_{22}\|_F + \frac{2KM\epsilon_0}{1-(M+K)\epsilon_1}]\|X_k\|_F +$$

$$+\frac{(M+K)\epsilon_1}{1-(M+K)\epsilon_1}\|A_{12}\|_F\|X_k\|_F^2 + \frac{4N\sqrt{MK}}{1-(N-1)\epsilon_1/2}\epsilon_0\}.$$

Thus,

$$\beta_1 = \frac{(K+3)\epsilon_1}{1-(K+5)\epsilon_1/2}\|\Delta_{11}\|_F + \frac{(M+3)\epsilon_1}{1-(M+5)\epsilon_1/2}\|\Delta_{22}\|_F + \frac{2KM\epsilon_0}{1-(M+K+3)\epsilon_1},$$

$$\beta_2 = \frac{(M+K+3)\epsilon_1}{1-(M+K+3)\epsilon_1}\|A_{12}\|_F, \quad \beta_3 = \frac{3\epsilon_1}{1-2\epsilon_1}\|A_{21}\|_F + \frac{4(M+K)\sqrt{MK}}{1-(M+K+5)\epsilon_1/2}\epsilon_0.$$

With the notation $F = \Delta_{22}X_k - X_k\Delta_{11} + A_{21} - X_kA_{12}X_k + \xi_1$ one has $X_{k+1} = -\mathcal{L}^{-1}(F + \xi_2) + \xi_3$, where $X_{k+1} = -(\mathcal{L}^{-1}F)_{mach}$. Here we use the fact that $\mathcal{L}$ is a diagonal matrix:

$$(\mathcal{L}^{-1}F)_{ij} = F_{ij}/(\Lambda_{1_i} - \Lambda_{2_j}),$$

$$(\Lambda_{1_i} - \Lambda_{2_j})_{mach} = \Lambda_{1_i}(1+t_1) - \Lambda_{2_j}(1+t_2), \quad |t_1| \le \epsilon_1, \quad |t_2| \le \epsilon_1.$$

We shall assume that $\tau > \epsilon_1(\|\Lambda_1\| + \|\Lambda_2\|)$. Then

$$X_{k+1} = -\frac{F_{ij}(1+t_3)}{\Lambda_{1_i}(1+t_1) - \Lambda_{2_j}(1+t_2)} + \xi_{ij}, \quad |t_3| \le \epsilon_1, \quad |\xi_{ij}| \le \epsilon_0,$$

$$(\xi_2)_{ij} = F_{ij}\left[\frac{(\Lambda_{1_i} - \Lambda_{2_j})(1+t_3)}{\Lambda_{1_i}(1+t_1) - \Lambda_{2_j}(1+t_2)} - 1\right], \quad \|\xi_3\|_F \le \epsilon_0\sqrt{MK} = \gamma_2,$$

$$\frac{(\Lambda_{1_i} - \Lambda_{2_j})(1+t_3)}{\Lambda_{1_i}(1+t_1) - \Lambda_{2_j}(1+t_2)} - 1 = \frac{\Lambda_{1_i}(t_3 - t_1) - \Lambda_{2_j}(t_3 - t_2)}{\Lambda_{1_i} - \Lambda_{2_j} + \Lambda_{1_i}t_1 - \Lambda_{2_j}t_2} \le$$

$$\le t_3 - (1+t_3)\frac{\Lambda_1 t_1 - \Lambda_2 t_2}{\Lambda_1 - \Lambda_2 + \Lambda_1 t_1 - \Lambda_2 t_2} \le \gamma_1 = \epsilon_1 + \frac{\epsilon_1(1+\epsilon_1)(\|\Lambda_1\| + \|\Lambda_2\|)}{\tau - \epsilon_1(\|\Lambda_1\| + \|\Lambda_2\|)}.$$

Note that for majority of modern computers (apart from CRAY), in particular, keeping the IEEE standard, $t_1 = t_2$. In the case we have $\gamma_1 = 2\epsilon_1/(1-\epsilon_1)$ which is essentially better than $\gamma_1 = \epsilon_1 + \frac{\epsilon_1(1+\epsilon_1)(\|\Lambda_1\|+\|\Lambda_2\|)}{\tau-\epsilon_1(\|\Lambda_1\|+\|\Lambda_2\|)}$.

5 Block deflation with higher precision

The spectral decomposition of the matrix A_1 is computed sequentially, block by block. Consider one step of the reduction. Having fixed τ one can represent the matrix A_1 in a block form:

$$A_1 = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

which is suitable for iterative solution of the Riccati equation

$$A_{22}R - RA_{11} + A_{21} - RA_{12}R = 0. \tag{17}$$

Choice of τ must be based on the results of section 3.

Suppose that $A_{ii} = \Lambda_i + \Delta_{ii}$ with the diagonal matrix Λ_i and small $\|\Delta_{ii}\|$, $i = 1, 2$. Assume that $\lambda_{\max}(\Lambda_1) - \lambda_{\min}(\Lambda_1) \leq \tau_1$ with small τ_1 (for example, it is easy to provide $\tau_1 < \tau(k-1)$, where k is the order of the matrix A_{11}) that is Λ_1 has the diagonals which are close to each other as much as possible.

One has $Q^T A_1 Q = \begin{pmatrix} \tilde{A}_{11} & 0 \\ 0 & \tilde{A}_{22} \end{pmatrix}$, where

$$Q = \begin{pmatrix} I & -R^T \\ R & I \end{pmatrix} \begin{bmatrix} (I + R^T R)^{-1/2} & 0 \\ 0 & (I + RR^T)^{-1/2} \end{bmatrix},$$

$$\tilde{A}_{11} = (I + R^T R)^{-1/2}[A_{11} + R^T A_{21} + A_{12}R + R^T A_{22}R](I + R^T R)^{-1/2},$$

$$\tilde{A}_{22} = (I + RR^T)^{-1/2}[A_{22} - RA_{12} - A_{21}R^T + RA_{11}R^T](I + RR^T)^{-1/2}.$$

The block deflation is the following process:

1) calculation of the matrix R from (17);

2) orthonormalization of the columns of $\begin{pmatrix} I \\ R \end{pmatrix}$ with higher precision. Resulting matrix is U_1;

3) orthonormalization of the columns of $\begin{pmatrix} -R^T \\ I \end{pmatrix}$ with higher precision. Resulting matrix is U_2;

4) calculation of $\tilde{A}_{11} = U_1^T A_1 U_1$ with higher precision;

5) calculation of $\tilde{A}_{22} = U_2^T A_1 U_2$ with higher precision.

Here for the orthonormalization we suggest to use the modified Gram-Schmidt orthogonalization which is assessed in section 6.

The matrix $\tilde{A}_{11}$ may have closed eigenvalues in the interval of length τ_1 approximately. This is true, for example, if $\|R\|$ is small enough. One can compute the matrix $\hat{\Delta}_{11} = \tilde{A}_{11} - \tilde{\lambda} I$ with small $\|\hat{\Delta}_{11}\|$ and $\tilde{\lambda} = (\lambda_{\max} + \lambda_{\min})/2$. Afterwards the spectral decomposition of $\hat{\Delta}_{11}$ with ordinary precision will be the spectral decomposition of $\tilde{A}_{11}$ with higher precision: $\tilde{A}_{11} = W^T \tilde{\Lambda}_1 W$. Here an orthogonal matrix W is stored in the factorized form i.e. by normals of the Householder transformations and by tangents (cotangents) of the rotation chains. Thus,

$$\begin{pmatrix} W & 0 \\ 0 & I \end{pmatrix}^T (U_1\ U_2)^T A_1 (U_1\ U_2) \begin{pmatrix} W & 0 \\ 0 & I \end{pmatrix} = \begin{pmatrix} \tilde{\Lambda}_1 & 0 \\ 0 & \tilde{A}_{22} \end{pmatrix}$$

is the spectral decomposition of A_1 with higher precision.

Repeating the above algorithm with the matrix $\tilde{A}_{22}$ and so on we shall come to the full spectral decomposition.

Let us turn now to the error analysis for one step of the deflation with higher precision. At first, remind that a solution to (17) may be calculated only with nonzero residual:

$$\|A_{22}R - RA_{11} + A_{21} - RA_{12}R\| \le \xi_0.$$

Therefore,

$$\left\| \begin{pmatrix} I & -R^T \\ R & I \end{pmatrix}^T \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} I & -R^T \\ R & I \end{pmatrix} - \right.$$

$$\left. - \begin{pmatrix} A_{11} + R^T A_{21} + A_{12}R + R^T A_{22}R & 0 \\ 0 & A_{22} - RA_{12} - A_{21}R^T + RA_{11}R^T \end{pmatrix} \right\| \le \xi_0.$$

Due to the rounding errors in the modified Gram-Schmidt algorithm:

$$\begin{pmatrix} I \\ R \end{pmatrix} = U_1 X_1 + \phi_1, \ \|\phi_1\| \le \xi_1, \ \|U_1^T U_1 - I\| \le \eta_1,$$

$$\begin{pmatrix} -R^T \\ I \end{pmatrix} = U_2 X_2 + \phi_2, \ \|\phi_2\| \le \xi_2, \ \|U_2^T U_2 - I\| \le \eta_2.$$

Hence

$$U^T U - I = (U_1\ U_2)^T (U_1\ U_2) - I = \left\{ [\begin{pmatrix} I \\ R \end{pmatrix} - \phi_1] X_1^{-1}, [\begin{pmatrix} -R^T \\ I \end{pmatrix} - \phi_2] X_2^{-1} \right\}^T \times$$

$$\times \left\{ [\begin{pmatrix} I \\ R \end{pmatrix} - \phi_1] X_1^{-1}, [\begin{pmatrix} -R^T \\ I \end{pmatrix} - \phi_2] X_2^{-1} \right\} - I =$$

$$= \begin{pmatrix} U_1^T U_1 - I & X_1^{-T} [\begin{pmatrix} I \\ R \end{pmatrix} - \phi_1]^T [\begin{pmatrix} -R^T \\ I \end{pmatrix} - \phi_2] X_2^{-1} \\ X_2^{-T} [\begin{pmatrix} -R^T \\ I \end{pmatrix} - \phi_2]^T [\begin{pmatrix} I \\ R \end{pmatrix} - \phi_1] X_1^{-1} & U_2^T U_2 - I \end{pmatrix},$$

$$\|U^T U - I\| \le \max(\eta_1, \eta_2) + \|X_2^{-T} [\begin{pmatrix} -R^T \\ I \end{pmatrix} - \phi_2]^T [\begin{pmatrix} I \\ R \end{pmatrix} - \phi_1] X_1^{-1}\| \le$$

$$\le \max(\eta_1, \eta_2) + \|X_2^{-1}\| \|\phi_2\| \|U_1\| + \|U_2\| \|\phi_1\| \|X_1^{-1}\| + \|\phi_1\| \|\phi_2\| \|X^{-1}\| \|X_2^{-1}\| \le$$

$$\le \max(\eta_1, \eta_2) + \frac{\sqrt{1+\eta_2}}{1-\xi_2} \xi_2 \sqrt{1+\eta_1} + \sqrt{1+\eta_2} \xi_1 \frac{\sqrt{1+\eta_1}}{1-\xi_1} + \xi_1 \xi_2 \frac{\sqrt{(1+\eta_1)(1+\eta_2)}}{(1-\xi_1)(1-\xi_2)} \le$$

$$\le \max(\eta_1, \eta_2) + \frac{\sqrt{(1+\eta_1)(1+\eta_2)}}{(1-\xi_1)(1-\xi_2)} [\xi_1 + \xi_2 - \xi_1 \xi_2].$$

Since

$$(U_1\ U_2)^T \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} (U_1\ U_2) - \begin{pmatrix} U_1^T A_1 U_1 & 0 \\ 0 & U_2^T A_1 U_2 \end{pmatrix} = \begin{pmatrix} 0 & U_1^T A_1 U_2 \\ U_2^T A_1 U_1 & 0 \end{pmatrix},$$

then

$$\|U^T A_1 U - \begin{pmatrix} U_1^T A_1 U_1 & 0 \\ 0 & U_2^T A_1 U_2 \end{pmatrix}\| \le \|U_2^T A_1 U_1\| =$$

$$= \|X_2^{-T} [\begin{pmatrix} -R^T \\ I \end{pmatrix} - \phi_2]^T A_1 [\begin{pmatrix} I \\ R \end{pmatrix} - \phi_1] X_1^{-1}\| \le \|X_2^{-1}\| \|\xi_0\| \|X_1^{-1}\| +$$

$$+\|U_2\|\|A_1\|\|\phi_1\|\|X_1^{-1}\| + \|X_2^{-1}\|\|\phi_2\|\|A_1\|\|U_1\| + \|X_2^{-1}\|\|\phi_2\|\|A_1\|\|\phi_1\|\|X_1^{-1}\| \le$$

$$\le \frac{\sqrt{(1+\eta_1)(1+\eta_2)}}{(1-\xi_1)(1-\xi_2)}\xi_0 + \sqrt{1+\eta_2}\|A_1\|\xi_1\frac{\sqrt{1+\eta_1}}{1-\xi_1} + \frac{\sqrt{1+\eta_2}}{1-\xi_2}\xi_2\|A_1\|\sqrt{1+\eta_1} +$$

$$+\frac{\sqrt{(1+\eta_1)(1+\eta_2)}}{(1-\xi_1)(1-\xi_2)}\xi_1\xi_2\|A_1\| \le \frac{\sqrt{(1+\eta_1)(1+\eta_2)}}{(1-\xi_1)(1-\xi_2)}[\xi_0 + \|A_1\|(\xi_1+\xi_2-\xi_1\xi_2)].$$

Remind once again that the diagonalization of $\tilde{A}_{11} = U_1^T A_1 U_1$ by means of diagonalization of $\hat{\Delta}_{11}$ with ordinary precision is valid if and only if $\|R\|$ is quite small. Otherwise, one has to subdivide $\tilde{A}_{11}$ into smaller blocks and to repeat the whole procedure starting from solution of the Riccati equations.

6 Modified Gram-Schmidt orthogonalization

In this section we assess the error analysis from [1] in order to take into account the underflow rounding errors.

Let A be a given $m \times n$-matrix of rank n, $m \ge n$. In the modified Gram-Schmidt process a sequence of matrices,

$$A = A^{(1)},\ A^{(2)},\ \ldots, A^{(n+1)} = Q,$$

is calculated. The matrix $A^{(k+1)}$ is computed from $A^{(k)} = (q_1, \ldots, q_{k-1}, a_k^{(k)}, \ldots, a_n^{(k)})$ by means of the following scheme:

$$r_{kk} = \|a_k^{(k)}\|, \quad q_k = a_k^{(k)}/r_{kk},$$
$$r_{kj} = (a_j^{(k)}, q_k), \quad a_j^{(k+1)} = a_j^{(k)} - r_{kj}q_k, \quad k+1 \le j \le n.$$

6.1 Errors in an elementary projection

Consider, at first, calculation of q_k. In order to compute q_k the following algorithm is used:

a) compute $v = a_k^{(k)}/\rho$ with $\rho = \max_i |(a_k^{(k)})_i|$;

b) compute $q_k = v/\sqrt{(v,v)}$.

The standard error analysis yields

$$\|v_{mach} - v\| \le \epsilon_1\|\frac{1}{\rho}a_{k_{mach}}^{(k)}\| + \epsilon_0\sqrt{m} \le (\epsilon_1 + \epsilon_0\sqrt{m})\|\frac{1}{\rho}a_{k_{mach}}^{(k)}\| = (\epsilon_1 + \epsilon_0\sqrt{m})\|v\|,$$

$$|(v,v)_{mach} - (v,v)| \le \frac{m\epsilon_1}{1-(m-1)\epsilon_1/2}\|v_{mach}\|^2 + \frac{(2m-1)\epsilon_0}{1-(m-1)\epsilon_1} + |\|v_{mach}\|^2 - \|v\|^2| \le$$

$$\le \frac{m\epsilon_1}{1-(m-1)\epsilon_1/2}\|v\|^2 + \frac{2\|v\|\|v_{mach}-v\|}{1-m\epsilon_1} + \frac{\|v_{mach}-v\|^2}{1-m\epsilon_1} + \frac{(2m-1)\epsilon_0}{1-(m-1)\epsilon_1} \le$$

$$\le \left[\frac{m\epsilon_1}{1-(m-1)\epsilon_1/2} + \frac{2(\epsilon_1+\epsilon_0\sqrt{m})}{1-m\epsilon_1} + \frac{(\epsilon_1+\epsilon_0\sqrt{m})^2}{1-m\epsilon_1} + \frac{(2m-1)\epsilon_0}{1-(m-1)\epsilon_1}\right]\|v\|^2 \le$$

$$\le \frac{(m+2)\epsilon_1 + (2\sqrt{m}+2m-1)\epsilon_0}{1-(m+1)\epsilon_1 - \epsilon_0\sqrt{m}}\|v\|^2,$$

$$|\sqrt{(v,v)_{mach}} - \sqrt{v,v)}| \le \|v\|\left[(1+\epsilon_1)\sqrt{1 + \frac{(m+2)\epsilon_1 + (2m+2\sqrt{m}-1)\epsilon_0}{1-(m+1)\epsilon_1 - \epsilon_0\sqrt{m}}} - 1\right] \le$$

$$\leq \frac{(m/2+2)\epsilon_1 + (m+\sqrt{m}-1/2)\epsilon_0}{1-(m+2)\epsilon_1 - \epsilon_0\sqrt{m}}\|v\|,$$

$$\|(q_k)_{mach} - q_k\| \leq \epsilon_1 \|\frac{v_{mach}}{\sqrt{(v,v)_{mach}}}\| + \epsilon_0\sqrt{m} \leq \epsilon_1 \frac{1+\epsilon_1+\epsilon_0\sqrt{m}}{1-\frac{(m/2+2)\epsilon_1+(m+\sqrt{m}-1/2)\epsilon_0}{1-(m+2)\epsilon_1-\epsilon_0\sqrt{m}}} + \epsilon_0\sqrt{m} \leq$$

$$\leq \frac{\epsilon_1 + \epsilon_0\sqrt{m}}{1-(3m/2+4)\epsilon_1 - (m+2\sqrt{m})\epsilon_0}. \tag{18}$$

A simple corollary from the previous estimate:

$$a^{(k)}_{k_{mach}} = (q_k)_{mach} r_{kk} + \xi^{(k)}_k, \quad \|\xi^{(k)}_k\| \leq \frac{\epsilon_1 + \epsilon_0\sqrt{m}}{1-(3m/2+4)\epsilon_1 - (m+2\sqrt{m})\epsilon_0}\|a^{(k)}_{k_{mach}}\|. \tag{19}$$

6.2 Errors in the decomposition

Now turn to error analysis when calculating $a^{(k+1)}_j$:

$$|(r_{kj})_{mach} - r_{kj}| \leq \frac{m\epsilon_1}{1-(m-1)\epsilon_1/2}\|(q_k)_{mach}\|\|a^{(k)}_{j_{mach}}\| + \frac{(2m-1)\epsilon_0}{1-(m-1)\epsilon_1},$$

$$|(r_{kj}q_k)_{mach} - r_{kj}q_{k_{mach}}| \leq \left(\epsilon_1|r_{kj}| + \frac{m\epsilon_1}{1-(m+1)\epsilon_1/2}\|q_{k_{mach}}\|\|a^{(k)}_{j_{mach}}\| + \right.$$

$$\left. + \frac{(2m-1)\epsilon_0}{1-m\epsilon_1}\right)\|(q_k)_{mach}\| + \epsilon_0\sqrt{m} \leq$$

$$\leq \frac{(m+1)\epsilon_1}{1-(m+1)\epsilon_1/2}\|a^{(k)}_{j_{mach}}\|\|(q_k)_{mach}\|^2 + \frac{(2m-1)\epsilon_0}{1-m\epsilon_1}\|(q_k)_{mach}\| + \epsilon_0\sqrt{m} \leq$$

$$\leq \frac{(m+1)\epsilon_1\|a^{(k)}_{j_{mach}}\|}{1-(7m+17)\epsilon_1/2 - (2m+4\sqrt{m})\epsilon_0} + \frac{(2m+\sqrt{m}-1)\epsilon_0}{1-(5m/2+4)\epsilon_1 - (m+2\sqrt{m})\epsilon_0}.$$

Thus,

$$a^{(k)}_{j_{mach}} = a^{(k+1)}_{j_{mach}} + r_{kj}(q_k)_{mach} + \xi^{(k)}_j, \tag{20}$$

$$\|\xi^{(k)}_j\| \leq \epsilon_1\|a^{(k)}_{j_{mach}}\| + \epsilon_1\|a^{(k)}_{j_{mach}}\|\|(q_k)_{mach}\|^2 + \epsilon_0\sqrt{m} + \frac{|(r_{kj}q_k)_{mach} - r_{kj}(q_k)_{mach}|}{1-\epsilon_1} \leq$$

$$\leq \frac{(m+3)\epsilon_1}{1-(7m+19)\epsilon_1/2 - (2m+4\sqrt{m})\epsilon_0}\|a^{(k)}_{j_{mach}}\| + \frac{(2m+2\sqrt{m}-1)\epsilon_0}{1-(5m+10)\epsilon_1/2 - (m+2\sqrt{m})\epsilon_0}.$$

Henceforth we shall use the following notations: $(q_k)_{mach} = \tilde{q}_k$, $a^{(k)}_{j_{mach}} = \tilde{a}^{(k)}_j$.
From

$$\|\tilde{a}^{(k+1)}_j\| \leq \|\tilde{a}^{(k)}_j - r_{kj}\tilde{q}_k\| + \|\xi^{(k)}_j\| = \|(I - \tilde{q}_k\tilde{q}_k^T)\tilde{a}^{(k)}_j\| + \|\xi^{(k)}_j\| \leq$$

$$\leq \|I - q_kq_k^T - (\tilde{q}_k - q_k)q_k^T - q_k(\tilde{q}_k - q_k)^T - (\tilde{q}_k - q)(\tilde{q}_k - q_k)^T\|\|\tilde{a}^{(k)}_j\| + \|\xi^{(k)}_j\| \leq$$

$$\leq \left(1 + \frac{(m+3)\epsilon_1}{1-(7m+19)\epsilon_1/2 - (2m+4\sqrt{m})\epsilon_0} + \right.$$

$$\left. + \frac{2(\epsilon_1+\epsilon_0\sqrt{m})}{1-(3m/2+4)\epsilon_1 - (m+2\sqrt{m})\epsilon_0 - \epsilon_1/2 - \sqrt{m}\epsilon_0/2}\right)\|\tilde{a}^{(k)}_j\| +$$

$$+ \frac{(2m+2\sqrt{m}-1)\epsilon_0}{1-(5m/2+5)\epsilon_1 - (m+2\sqrt{m})\epsilon_0} \leq \frac{\|\tilde{a}^{(k)}_j\|}{1-(7m+19)\epsilon_1/2 - (2m+4\sqrt{m})\epsilon_0} +$$

$$+\frac{(2m+2\sqrt{m}-1)\epsilon_0}{1-(5m/2+5)\epsilon_1-(m+2\sqrt{m})\epsilon_0}=p\|\tilde{a}_j^{(k)}\|+q$$

it follows that

$$\|\tilde{a}_j^{(k)}\|\le p^{k-1}\|\tilde{a}_j^{(1)}\|+\frac{p^{k-1}-1}{p-1}q\le\|\tilde{a}_j^{(1)}\|\frac{1}{1-[(7m+19)\epsilon_1/2+(2m+4\sqrt{m})\epsilon_0](k-1)}+$$

$$+\frac{(k-1)(2m+2\sqrt{m})\epsilon_0}{1-[(7m+19)\epsilon_1/2+(2m+4\sqrt{m})\epsilon_0](k-1)} \tag{21}$$

and

$$\tilde{a}_k^{(k)}=\tilde{q}_k r_{kk}+\xi_k^{(k)},\quad \|\xi_k^{(k)}\|\le\frac{(\epsilon_1+\epsilon_0\sqrt{m})[\|\tilde{a}_k^{(1)}\|+(k-1)(2m+2\sqrt{m})\epsilon_0]}{1-[(7m+19)\epsilon_1/2+(2m+4\sqrt{m})\epsilon_0]k}, \tag{22}$$

$$\tilde{a}_j^{(k)}=\tilde{a}_j^{(k+1)}+\tilde{q}_k r_{kj}+\xi_j^{(k)},\quad \|\xi_j^{(k)}\|\le\frac{(m+3)\epsilon_1}{1-[(7m+19)\epsilon_1/2+(2m+4\sqrt{m})\epsilon_0]k}\|\tilde{a}_j^{(1)}\|+ \tag{23}$$

$$+\frac{(2m+2\sqrt{m})\epsilon_0}{1-[(7m+19)\epsilon_1/2+(2m+4\sqrt{m})\epsilon_0]k}.$$

Since $\tilde{a}_j^{(1)}=\sum_{i=1}^{j}\tilde{q}_i r_{ij}+\sum_{i=1}^{j}\xi_j^{(i)}$ then

$$\|\tilde{a}_j^{(1)}-\sum_{i=1}^{j}\tilde{q}_i r_{ij}\|\le\|\tilde{a}_j^{(1)}\|\frac{[(m+3)(j-1)+1]\epsilon_1+\sqrt{m}\epsilon_0}{1-[(7m+19)\epsilon_1/2+(2m+4\sqrt{m})\epsilon_0]j}+$$

$$+\frac{(1+\epsilon_1+\epsilon_0\sqrt{m})(2m+2\sqrt{m})(j-1)\epsilon_0}{1-[(7m+19)\epsilon_1/2+(2m+4\sqrt{m})\epsilon_0]j},$$

$$\|A-\tilde{Q}R\|_F\le\alpha\|A\|_F+\overline{\alpha}, \tag{24}$$

where

$$\tilde{Q}=[\tilde{q}_1,\tilde{q}_2,\ldots,\tilde{q}_n],\quad R=\begin{pmatrix} r_{11} & r_{12} & \ldots & r_{1n}\\ & r_{22} & \ldots & r_{2n}\\ & & \ldots & \\ \mathrm{O} & & & r_{nn}\end{pmatrix},$$

$$\alpha=\frac{[(m+3)(n-1)+1]\epsilon_1+\sqrt{m}\epsilon_0}{1-[(7m+19)\epsilon_1/2+(2m+4\sqrt{m})\epsilon_0]n},$$

$$\overline{\alpha}=\frac{(1+\epsilon_1+\epsilon_0\sqrt{m})(2m+2\sqrt{m})(n-1)\epsilon_0}{1-[(7m+19)\epsilon_1/2+(2m+4\sqrt{m})\epsilon_0]n},$$

6.3 Orthogonality of the computed vectors

Here we estimate deviation from orthonormality for the vectors $\tilde{q}_j$. This can be done by means of computation of $\|\tilde{Q}^T\tilde{Q}-I\|$. As $\tilde{a}_j^{(l)}=\sum_{i=l}^{j}\tilde{q}_i r_{ij}+\sum_{i=l}^{j}\xi_j^{(i)}$, then the identities

$$\tilde{q}_{l-1}^T\tilde{a}_j^{(l)}=\tilde{q}_{l-1}^T(\tilde{a}_j^{(l-1)}-r_{l-1,j}\tilde{q}_{l-1}-\xi_j^{(l-1)})=\tilde{q}_{l-1}^T(I-\tilde{q}_{l-1}\tilde{q}_{l-1}^T)\tilde{a}_j^{(l-1)}-\tilde{q}_{l-1}^T\xi_j^{(l-1)},$$

$$\tilde{q}_{l-1}^T\tilde{a}_j^{(l)}=\sum_{i=l}^{j}(\tilde{q}_{l-1}^T\tilde{q}_i)r_{ij}+\tilde{q}_{l-1}^T\sum_{i=l}^{j}\xi_j^{(i)}$$

imply that

$$\sum_{i=l}^{j}(\tilde{q}_{l-1}^T\tilde{q}_i)r_{ij}=\tilde{q}_{l-1}^T(I-\tilde{q}_{l-1}^T\tilde{q}_{l-1})\tilde{a}_j^{(l-1)}-\tilde{q}_{l-1}^T\sum_{i=l-1}^{j}\xi_j^{(i)},\quad 2\le l\le j\le n.$$

Let us introduce a $n \times n$-matrix U which is upper strict triangular and $U_{rs} = (\tilde{q}_r, \tilde{q}_s)$, $r < s$. Then $\sum_{i=l}^{j} (\tilde{q}_{l-1}^T \tilde{q}_i) r_{ij} = (UR)_{l-1,j}$, where the matrix UR is also upper strict triangular. It follows that

$$(UR)_{l-1,j} = \tilde{q}_{l-1}^T \left[(I - \tilde{q}_{l-1}^T \tilde{q}_{l-1}) \tilde{a}_j^{(l-1)} - \sum_{i=l-1}^{j} \xi_j^{(i)} \right], \quad 2 \le l \le j \le n,$$

$$|(UR)_{l-1,j}| \le \frac{2(\epsilon_1 + \epsilon_0\sqrt{m})}{1 - (3m/2+6)\epsilon_1 - (m + 4\sqrt{m})\epsilon_0} \|\tilde{a}_j^{(l-1)}\| +$$

$$+ \|\tilde{a}_j^{(l-1)}\| \frac{[(m+3)(j-l+1)+1]\epsilon_1 + \sqrt{m}\epsilon_0}{1 - [(7m+19)\epsilon_1/2 + (2m+4\sqrt{m})\epsilon_0](j-l+2)} +$$

$$+ \frac{(1 + \epsilon_1 + \epsilon_0\sqrt{m})(2m+2\sqrt{m})(j-l+1)\epsilon_0}{1 - [(7m+19)\epsilon_1/2 + (2m+4\sqrt{m})\epsilon_0](j-l+2)} \le$$

$$\le \|\tilde{a}_j^{(l-1)}\| \frac{[(m+3)(j-l+1)+3]\epsilon_1 + 3\sqrt{m}\epsilon_0}{1 - [(7m+19)\epsilon_1/2 + (2m+4\sqrt{m})\epsilon_0](j-l+2)} +$$

$$+ \frac{(1 + \epsilon_1 + \epsilon_0\sqrt{m})(2m+2\sqrt{m})(j-l+1)\epsilon_0}{1 - [(7m+19)\epsilon_1/2 + (2m+4\sqrt{m})\epsilon_0](j-l+2)} \le$$

$$\le \|\tilde{a}_j^{(1)}\| \frac{[(m+3)(j-l+1)+3]\epsilon_1 + 3\sqrt{m}\epsilon_0}{1 - [(7m+19)\epsilon_1/2 + (2m+4\sqrt{m})\epsilon_0]j} +$$

$$+ \frac{(2m+2\sqrt{m})\epsilon_0}{1 - [(7m+19)\epsilon_1/2 + (2m+4\sqrt{m})\epsilon_0]j} \times$$

$$\times \{(1 + \epsilon_1 + \epsilon_0\sqrt{m})(j-l+1)(1 - [(7m+19)\epsilon_1/2 + (2m+4\sqrt{m})\epsilon_0][l-2]) +$$

$$+ (l-2)([(m+3)(j-l+1)+3]\epsilon_1 + 3\sqrt{m}\epsilon_0)\} \le$$

$$\le \|\tilde{a}_j^{(1)}\| \frac{[(m+3)(j-l+1)+3]\epsilon_1 + 3\sqrt{m}\epsilon_0}{1 - [(7m+19)\epsilon_1/2 + (2m+4\sqrt{m})\epsilon_0]j} +$$

$$+ \frac{(1 + \epsilon_1 + \epsilon_0\sqrt{m})(j-l+1)(2m+2\sqrt{m})\epsilon_0}{1 - [(7m+19)\epsilon_1/2 + (2m+4\sqrt{m})\epsilon_0]j},$$

$$\|(UR)_j\| \le \frac{\|\tilde{a}_j^{(1)}\| \left[((m+3)\sqrt{j(j-1)(2j-1)/6} + 3\sqrt{j-1})\epsilon_1 + 3\sqrt{m}\sqrt{j-1}\epsilon_0 \right]}{1 - [(7m+19)\epsilon_1/2 + (2m+4\sqrt{m})\epsilon_0]j} +$$

$$+ \frac{(1 + \epsilon_1 + \epsilon_0\sqrt{m})(2m+2\sqrt{m})\sqrt{j(j-1)(2j-1)/6}\epsilon_0}{1 - [(7m+19)\epsilon_1/2 + (2m+4\sqrt{m})\epsilon_0]j}.$$

As a result,

$$\|UR\|_F \le \beta \|A\|_F + \overline{\beta}, \tag{25}$$

where

$$\beta = \frac{\sqrt{n-1}\{[(m+3)\sqrt{n(2n-1)/6} + 3]\epsilon_1 + 3\sqrt{m}\epsilon_0\}}{1 - [(7m+19)\epsilon_1/2 + (2m+4\sqrt{m})\epsilon_0]n},$$

$$\overline{\beta} = \frac{(1 + \epsilon_1 + \epsilon_0\sqrt{m})(2m+2\sqrt{m})\sqrt{n^2(n^2-1)/12}\epsilon_0}{1 - [(7m+19)\epsilon_1/2 + (2m+4\sqrt{m})\epsilon_0]n}.$$

One can easily show that $\tilde{Q}^T\tilde{Q} = D + U^T + U$ with the diagonal matrix D for which the following estimate is valid:

$$\|D - I\| \le \gamma = \frac{2(\epsilon_1 + \epsilon_0\sqrt{m})}{1 - (3m/2+5)\epsilon_1 - (m + 3\sqrt{m})\epsilon_0}.$$

Since

$$\begin{cases} A = \tilde{Q}R + \Delta_1, \quad \|\Delta_1\|_F \le \alpha\|A\|_F + \overline{\alpha}, \\ \tilde{Q}^T\tilde{Q} = D + U^T + U, \quad \|D - I\| \le \gamma, \\ UR = \Delta_2, \quad \|\Delta_2\|_F \le \beta\|A\|_F + \overline{\beta}, \end{cases} \tag{26}$$

then

$$A^TA = R^T\tilde{Q}^T\tilde{Q}R + \Delta_1^T(A - \Delta_1) + (A - \Delta_1)^T\Delta_1 + \Delta_1^T\Delta_1 =$$
$$= R^TR + R^T(D - I)R + R^TU^TR + R^TUR + \Delta_1^TA + A^T\Delta_1 - \Delta_1\Delta_1,$$
$$R^TR = (A - \Delta_1)^T(A - \Delta_1) - R^T(D - I)R - \Delta_2^TR - R^T\Delta_2.$$

It follows that

$$\sigma_{\min}^2(R) \ge (\sigma_{\min}(A) - \|\Delta_1\|)^2 - \gamma\|R\|^2 - 2\|\Delta_2\|\|R\|,$$
$$\|R\|^2 \le (\|A\| + \|\Delta_1\|)^2 + \gamma\|R\|^2 + 2\|\Delta_2\|\|R\|.$$

From the second equation one can get the following bound:

$$\|R\| \le r = \frac{\|\Delta_2\| + \sqrt{\|\Delta_2\|^2 + (1 - \gamma)(\|A\| + \|\Delta_1\|)^2}}{1 - \gamma}.$$

It holds that

$$r^2 - (\|A\| + \|\Delta_1\|)^2 \le [2\|\Delta_2\|^2 + (1 - \gamma)(\|A\| + \|\Delta_1\|)^2 +$$
$$+2\|\Delta_2\|\sqrt{\|\Delta_2\|^2 + (1 - \gamma)(\|A\| + \|\Delta_1\|)^2}]/(1 - \gamma)^2 - (\|A\| + \|\Delta_1\|)^2 \le$$
$$\le \frac{\gamma}{1 - \gamma}(\|A\| + \|\Delta_1\|)^2 + \frac{2\|\Delta_2\|(\|\Delta_2\| + \sqrt{\|\Delta_2\|^2 + (1 - \gamma)(\|A\| + \|\Delta_1\|)^2})}{(1 - \gamma)^2} \le$$
$$\le \frac{\gamma}{1 - \gamma}(\|A\| + \|\Delta_1\|)^2 + \frac{2\|\Delta_2\|(2\|\Delta_2\| + \|A\| + \|\Delta_1\|)}{(1 - \gamma)^2},$$
$$\sigma_{\min}^2(R) \ge (\sigma_{\min}(A) - \|\Delta_1)^2 - [r^2 - (\|A\| + \|\Delta_1\|)^2] \ge (\sigma_{\min}(A) - \|\Delta_1\|)^2 -$$
$$-\frac{\gamma}{1 - \gamma}(\|A\| + \|\Delta_1\|)^2 - \frac{2\|\Delta_2\|(2\|\Delta_2\| + \|A\| + \|\Delta_1\|)}{(1 - \gamma)^2},$$
$$\|\tilde{Q}^T\tilde{Q} - I\| \le \|D - I\| + 2\|\Delta_2\|\|R^{-1}\| \le$$
$$\le \gamma + \frac{2\|\Delta_2\|}{\sqrt{(\sigma_{\min}(A) - \|\Delta_1\|)^2 - \frac{\gamma}{1-\gamma}(\|A\| + \|\Delta_1\|)^2 - \frac{2\|\Delta_2\|(2\|\Delta_2\| + \|A\| + \|\Delta_1\|)}{(1-\gamma)^2}}} \le$$
$$\le \chi = \gamma + \frac{2\|\Delta_2\|}{\sigma_{\min}(A) - \|\Delta_1\|} \times \tag{27}$$
$$\times \frac{1}{\left[1 - \frac{\gamma}{1-\gamma}\left(\frac{\|A\| + \|\Delta_1\|}{\sigma_{\min}(A) - \|\Delta_1\|}\right)^2 - \frac{2\|\Delta_2\|(\|A\| + \|\Delta_1\| + 2\|\Delta_2\|)}{(1-\gamma)^2(\sigma_{\min}(A) - \|\Delta_1\|)^2}\right]}.$$

This latter estimate is valid if and only if

$$\frac{\gamma}{1 - \gamma}\left(\frac{\|A\| + \|\Delta_1\|}{\sigma_{\min}(A) - \|\Delta_1\|}\right)^2 - \frac{2\|\Delta_2\|(\|A\| + \|\Delta_1\| + 2\|\Delta_2\|)}{(1 - \gamma)^2(\sigma_{\min}(A) - \|\Delta_1\|)^2} < 1. \tag{28}$$

Consider the estimates (27), (28) when $\|A\| \approx 1$, $\epsilon_0 \ll \epsilon_1$, $mn\epsilon_1 \ll 1$. One has

$$\|\Delta_1\| \le\approx [(m + 3)(n - 1) + 1]\sqrt{n}\epsilon_1\|A\|,$$

$$\|\Delta_2\| \leq\approx \sqrt{(n-1)n}[(m+3)\sqrt{n(2n-1)/6}+3]\epsilon_1\|A\|,$$

$$\gamma \leq\approx 2\epsilon_1.$$

Introduce the notation $\mu = \|A\|\|A^{-1}\|$. Then the inequality (28) implies that

$$\frac{2\epsilon_1\mu^2}{1-2\epsilon_1}\left\{\frac{1+\sqrt{n}[(m+3)(n-1)+1]\epsilon_1}{1-\sqrt{n}[(m+3)(n-1)+1]\epsilon_1\mu}\right\}^2 + \frac{2\epsilon_1\sqrt{(n-1)n}[(m+3)\sqrt{n(2n-1)/6}+3]}{(1-2\epsilon_1-\sqrt{n}[(m+3)(n-1)+1]\epsilon_1\mu)^2}\times$$

$$\times\{1+[(m+3)(n-1)+1]\sqrt{n}\epsilon_1 + 2\sqrt{(n-1)n}[(m+3)\sqrt{n(2n-1)/6}+3]\epsilon_1\}\mu^2 \leq$$

$$\leq 2\epsilon_1\mu^2\{1+\sqrt{(n-1)n}[(m+3)\sqrt{n(2n-1)/6}+3]\}/\{1-4\epsilon_1-$$

$$-2\sqrt{n}[(m+3)(n-1)+1]\epsilon_1(\mu+1) - 2\sqrt{(n-1)n}[(m+3)\sqrt{n(2n-1)/6}+3]\epsilon_1\} \approx< 1.$$

One can also show that

$$\chi \leq\approx 2\epsilon_1 + \frac{2\epsilon_1\mu\sqrt{(n-1)n}[(m+3)\sqrt{n(2n-1)/6}+3]}{1-2\epsilon_1\mu^2\{1+\sqrt{(n-1)n}[(m+3)\sqrt{n(2n-1)/6}+3]\}}.$$

7 Refinement of invariant subspaces

Sometimes one has to calculate with high precision only a few eigenvectors which give a basis of an invariant subspace of a symmetric matrix. If there is roughly computed full spectral decomposition of this symmetric matrix then we can suggest another refinement procedure but also based upon the Riccati equations. It might be less costly for double precision operations.

Given a symmetric $N \times N$-matrix A. Assume that $x_1, x_2, \ldots, x_k$, $y_{k+1}, \ldots, y_N$ are approximate eigenvectors of A corresponding to approximate eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$, $\lambda_{k+1}, \ldots, \lambda_N$. We suppose that k is small enough. Let the columns of $N \times k$-matrix X be the vectors $x_1, x_2, \ldots, x_k$ and so is for $N \times (N-k)$-matrix Y and vectors $y_{k+1}, \ldots, y_N$. Introduce the following notations:

$$\Lambda_1 = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_k \end{pmatrix}, \quad \Lambda_2 = \begin{pmatrix} \lambda_{k+1} & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix}.$$

Then $\|A[X \vdots Y] - [X \vdots Y]\Lambda\| \leq \zeta$ with small ζ depending on the method of computation of the crude spectral decomposition.

We can try to look for better approximation of the invariant subspace defined by $x_1, x_2, \ldots, x_k$ in the form $\hat{X} = X + YR$ with $(N-k) \times k$-matrix R. The columns of $\hat{X}$ will give a basis for the subspace.

If $A_1 = (X \vdots Y)^{-1}A(X \vdots Y)$, where division on blocks conforms to dimensions of X and Y, then the matrix R must obey the equation

$$(-R \vdots I)A_1\begin{pmatrix} I \\ R \end{pmatrix} = 0. \tag{29}$$

In order to approximate the matrix A_1 we made use of the matrix

$$\tilde{A}_1 = \Lambda + (X \vdots Y)^T[A(X \vdots Y) - (X \vdots Y)\Lambda]$$

with the error $A_1 - \tilde{A}_1 = [(X \vdots Y)^{-1} - (X \vdots Y)^T][A(X \vdots Y) - (X \vdots Y)\Lambda]$ which is quite small when the matrix $(X \vdots Y)$ is almost orthogonal and the residual $A(X \vdots Y) - (X \vdots Y)\Lambda$ is small.

So one can replace A_1 in (29) by $\tilde{A}_1$:

$$(-R \vdots I)\{\Lambda + (X \vdots Y)^T[A(X \vdots Y) - (X \vdots Y)\Lambda]\} \begin{pmatrix} I \\ R \end{pmatrix} = 0. \tag{30}$$

In the iterative process (16) we have to calculate the matrix $F = A_{22}R - RA_{11} + A_{21} - RA_{12}R$ with high precision. With the notation $A_1 = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ equation (30) yields

$$F = (-R \vdots I)(X \vdots Y)^T[A(X \vdots Y) - (X \vdots Y)\Lambda] \begin{pmatrix} I \\ R \end{pmatrix}. \tag{31}$$

After some transformations

$$F = (Y^T - RX^T)[A(X + YR) - (X\Lambda_1 + Y\Lambda_2 R)] = (Y^T - RX^T)[(AX - X\Lambda_1) + (AYR - Y\Lambda_2 R)]. \tag{32}$$

How to calculate (32) with high precision ? We suggest the following steps:

1) Calculation of $F_X = AX - X\Lambda_1$ with higher precision;

2) Calculation of $F_Y = A(YR) - Y(\Lambda_2 R)$ with higher precision in the order prescribed by parentheses;

3) Calculation of $F = (Y^T - RX^T)(F_X + F_Y)$ with ordinary precision.

Error analysis confirms this procedure if $\|AX - X\Lambda_1\|$ and $\|AY - Y\Lambda_2\|$ are small enough.

The prescribed order of operations and general motivation for this approach are based upon the fact that higher precision multiplications $AX, X\Lambda_1, YR, A(YR), \Lambda_2 R, Y(\Lambda_2 R)$ are not very costly if k is sufficiently small. Assume that multiplication of two $M \times N$ and $N \times K$ matrices takes $2MNK$ operations, then the above multiplications cost $2N^2k + 2Nk^2 + 2N(N_k)k + 2N^2k + 2(N-k)^2k + 2N(N-k)k = 2(5N^2k - 3Nk^2 + k^3)$ higher precision operations. Maximum of the number of operations is reached when $k = N$ and equals to $6N^3$. Notice that the multiplications AX and $X\Lambda_1$ are calculated only once in (16) because they do not depend on R.

The rounding error analysis from previous sections can be modified for this new refinement. We do not provide it here as there is no difficulties to develop the theory analogous to that in sections 2 and 3.

Now we will discuss possible extension of the above technique for the case when the matrix A_1 is badly approximated by $\tilde{A}_1$. The idea of this extension is adopted from [2].

Consider the following iteration:

$$B_0 = \Lambda, \quad B_{k+1} = B_k + (X \vdots Y)^T[A(X \vdots Y) - (X \vdots Y)B_k]. \tag{33}$$

Since $A_1 = A_1 + (X \vdots Y)^T[A(X \vdots Y) - (X \vdots Y)A_1]$ then

$$B_{k+1} - A_1 = B_k - A_1 - (X \vdots Y)^T(X \vdots Y)(B_k - A_1) = [I - (X \vdots Y)^T(X \vdots Y)](B_k - A_1). \tag{34}$$

Therefore, if $0 < (X \vdots Y)^T(X \vdots Y) < I$ then the iteration (33) converges.

We can transform (33) in the following way:

$$B_{k+1} - \Lambda = (B_k - \Lambda) + (X \vdots Y)^T\{[A(X \vdots Y) - (X \vdots Y)\Lambda] - (X \vdots Y)(B_k - \Lambda)\}.$$

Denote $B_k - \Lambda$ by C_k, then

$$C_0 = 0, \quad C_{k+1} = C_k - (X \vdots Y)^T (X \vdots Y) C_k + (X \vdots Y)^T [A(X \vdots Y) - (X \vdots Y)\Lambda]. \tag{35}$$

The matrix $H_k = C_k \begin{pmatrix} I \\ R \end{pmatrix}$ satisfies the iteration

$$H_0 = 0, \quad H_{k+1} = H_k - (X \vdots Y)^T (X \vdots Y) H_k + (X \vdots Y)^T [A(X \vdots Y) - (X \vdots Y)\Lambda] \begin{pmatrix} I \\ R \end{pmatrix} \tag{36}$$

and converges to $H_\infty = (A_1 - \Lambda) \begin{pmatrix} I \\ R \end{pmatrix}$. Thus, the matrix F is equal to $(-R \vdots I) H_\infty$.

Hence, in order to calculate the matrix F being used in iteration (16) one can compute H_∞ from the iteration

$$H_0 = 0, \quad H_{k+1} = H_k - (X \vdots Y)^T (X \vdots Y) H_k + (X \vdots Y)^T [A(AX - X\Lambda_1) + (AYR - Y\Lambda_2 R)] \tag{37}$$

and then $F = (-R \vdots I) H_\infty$. Multiplications of matrices must be executed in the following order:

$$(X \vdots Y)^T [(X \vdots Y) H_k], \quad A[YR], \quad Y[\Lambda_2 R],$$

in order to have a low operation count for small k.

Consideration of the rounding errors and of convergence rate for the processes (37) and (16) is more complicated. In addition, note that one can possibly combine these two iterations in another way to obtain faster convergence.

Some final comments

In this paper the technique for solving some eigenvalue problems has been discussed. The approach developed here is intended to be less expensive than the calculation of the spectral decomposition with double precision entirely that is calculation of the Householder transformations and then the deflation of tridiagonal matrices with double precision.

Our method makes use of mainly the BLAS 3 operations [5] for higher precision. The BLAS 2 operations are used only in MGS orthonormalization (see [7] for the block version of MGS). Hence the method is well suitable for parallel computers.

In addition, our method is equipped with the strict mathematical theory for the convergence rate and for the rounding errors.

The technique used can be easily extended to computation of the singular value decomposition with high precision. For this one should apply orthonormalizations to the left and to the right singular vectors and then compute solution for the generalized Riccati equation. The Riccati equation for this case is outlined in [11]. There is no principal problems with development of iterative process to solve the generalized equation similar to that in section 3. The error analysis for the refinement of SVD is the same as exposed above.

Acknowledgement. I wish to thank Bernard Philippe for stimulating discussions.

References

[1] A. Björck. Solving linear least squares problems by Gram-Schmidt orthogonalization. *BIT*, 7:1–21, 1967.

[2] R. Choquet. Diploma Work, University Rennes I, France, 1991. (in French).

[3] J. Demmel. Three methods for refining estimates of invariant subspaces. *Computing*, 38:43–57, 1987.

[4] J. Demmel. Underflow and the reliability of numerical software. *SIAM J. Sci. Statist. Comput.*, 5:887–919, 1984.

[5] J.J. Dongarra, I.S. Duff, D.C. Sorensen, and H.A. Van der Vorst. *Solving Linear Systems on Vector and Shared Memory Computers.* SIAM Publications, Philadelphia, 1991.

[6] S.K. Godunov, A.G. Antonov, O.P. Kiriluk, and V.I. Kostin. *Guaranteed Accuracy of the Solution to Systems of Linear Equations in Euclidean Spaces.* Nauka, Novosibirsk, 1988. (in Russian).

[7] W. Jalby and B. Philippe. Stability analysis and improvement of the block Gram-Schmidt algorithm. *SIAM J. Sci. Statist. Comput.*, 12, 1991.

[8] R. Lohner. *Enclosing all Eigenvalues of Symmetric Matrices.* Technical Report DIAMOND Deliverable D3-1 part 3, Universität Karlsruhe, Germany, 1990.

[9] A.N. Malyshev. *Introduction to Numerical Linear Algebra.* Nauka, Novosibirsk, 1991. (in Russian).

[10] B. Philippe. *Perturbation de la Décomposition Spectrale d'une Matrice Hermitienne.* Technical Report 269, INRIA-IRISA, Centre de Rennes, IRISA, Campus de Beaulieu, 35042 Rennes Cédex, France, 1984.

[11] G.W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory.* Academic Press, San Diego, California, 1990.

LISTE DES PUBLICATIONS INTERNES IRISA 1992

PI 624 SIGNAL AS A MODEL FOR REAL-TIME AND HYBRID SYSTEMS
Albert BENVENISTE, Michel LE BORGNE, Paul LE GUERNIC
Janvier 1992, 22 pages.

PI 625 ON THE CENTRAL-LIMIT THEOREM FOR TRACKING ESTIMATORS WITH SMALL GAIN - INFINITE HORIZON CASE
Bernard DELYON, Anatoli JUDITSKY
Janvier 1992, 16 pages.

PI 626 A MONTE CARLO METHOD BASED ON ANTITHETIC VARIATES FOR NETWORK RELIABILITY COMPUTATIONS
Mohamed EL KHADIRI, Gerardo RUBINO
Janvier 1992, 28 pages.

PI 627 CONSTRAINED MULTISCALE MARKOV RANDOM FIELDS AND THE ANALYSIS OF VISUAL MOTION
Fabrice HEITZ, Patrick PEREZ, Patrick BOUTHEMY
Janvier 1992, 40 pages.

PI 628 ON ITERATIVE REFINEMENT FOR THE SPECTRAL DECOMPOSITION OF SYMMETRIC MATRICES
Alexander N. MALYSHEV
Janvier 1992, 26 pages.

PI 629 STRUCTURAL OPERATIONAL SPECIFICATIONS AND TRACE AUTOMATA
Eric BADOUEL, Philippe DARONDEAU
Janvier 1992, 36 pages.

PI 630 EREBUS, A DEBUGGER FOR ASYNCHRONOUS DISTRIBUTED COMPUTING SYSTEM
Michel HURFIN, Noël PLOUZEAU, Michel RAYNAL
Janvier 1992, 14 pages.

PI 631 PROTOCOLES SIMPLES POUR L'IMPLEMENTATION REPARTIE DES SEMAPHORES
Michel RAYNAL
Janvier 1992, 14 pages.

PI 632 L-STABLE PARALLEL ONE-BLOCK METHODS FOR ORDINARY DIFFERENTIAL EQUATIONS
Philippe CHARTIER, Bernard PHILIPPE
Janvier 1992, 28 pages.

PI 633 ON EFFICIENT CHARACTERIZING SOLUTIONS OF LINEAR DIOPHANTINE EQUATIONS AND ITS APPLICATION TO DATA DEPENDENCE ANALYSIS
Christine EISENBEIS, Olivier TEMAM, Harry WIJSHOFF
Janvier 1992, 22 pages.

PI 634 UN NOYAU DE SYSTEME REPARTI POUR LES APPLICATIONS GEREES PAR UN TEMPS VIRTUEL
Janvier 1992, 20 pages.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique

ISSN 0249 - 6399